AVIATION SECURITY RESOURCE MANAGEMENT:

A PSYCHOLOGICAL APPROACH TO PRE-EMPLOYMENT AND TRAINING PROCEDURES

Thesis presented to the Faculty of Arts
of
the University of Zurich
for the degree of Doctor of Philosophy

by
Diana Hardmeier
of
Zumikon / ZH

Accepted in the spring semester 2008 on the recommendation of
Prof. Dr. Wolfgang Marx and Prof. Dr. Mike Martin

Zentralstelle der Studentenschaft
Zurich 2008

CONTENT

1. SUMMARY

This thesis is about human factors in aviation security. It consists of three main sections that is employment of competent job applicants (PRE-EMPLOYMENT ASSESSMENT OF AVIATION SECURITY SCREENERS), training of employed screeners (TRAINING OF AVIATION SECURITY SCREENERS) and effects of age in X-ray screening (AGE EFFECTS IN AVIATION SECURITY SCREENING). Based on visual cognition research, psychophysical methods and psychological test theories, a reliable and valid pre-employment assessment has been developed which focuses on one of the most important tasks in aviation security, the X-ray screening of passenger bags. An X-ray screening test that is independent of knowledge was developed and evaluated. The X-Ray Object Recognition Test (X-Ray ORT) measures the ability to cope with the image-based factors bag complexity, superposition and viewpoint of threat items relatively independent of knowledge. Further, the influence of ability on detection performance was tested using a cognitive test battery which consists of 12 general cognition tests that best match the X-ray screening task. Structural equation modeling further revealed the three main factors ability, training and age. Therefore, the effect of training on detection performance was examined. Previous studies in the hold baggage screening (HBS) area showed a training effect for improvised explosive devices (IED) that are normally not seen at checkpoints and vary enormously in their visual appearance. The question whether this training effect can also be shown for other objects such as knives that are quite often seen at checkpoints in the cabin baggage screening (CBS) area is discussed in this study. Empirical results reveal that even in the CBS area where screeners encounter many different objects, training is essential to store all different types of prohibited items in the visual memory. Thus, working experience helps to familiarize with many different items, however it is not enough to acquire the full visual memory representation of threat items. Moreover, transfer effects could be observed implying that acquired detection skills can easily be generalized and used to detect new, but similar looking items. Regarding viewpoint effects, an increased detection performance could also be found for unusual views, however no significant interaction between viewpoint and training. Last, the effect of age on X-ray screening was investigated. One could assume that older screeners with more working experience should be able to compensate age related declines. Contrary to expectations overall results of a visual search task and threat image projection (TIP) data, that is detection performance on the job, provide evidence that age-related differences in job-specific fluid

performance even exist when persons are practiced in that ability and may use any strategy available to maximize their performance. Further analyses show that the age effect is different for tasks which require visual cognition abilities compared to tasks which require visual knowledge.

In summary, this thesis clearly showed the influence of visual cognition abilities for the X-ray screening task of aviation security screeners and therewith emphasizes the importance of a reliable and valid pre-employment assessment procedure. Further, working experience alone is not enough to familiarize with as many different threat items as needed to store all different types of prohibited items in the visual memory. A significant increase in detection performance due to an individually adaptive training system could be revealed. In addition, relatively large age effects could also be found for experienced aviation security screeners.

2. OUTLINE

For years terrorist attacks have presented a constant threat to civil aviation and since 9/11 a new dimension of threats poses an additional challenge to security measures. Thus, the importance of aviation security clearly increased during the last years. To improve and facilitate security processes, large investments into technology were made. Despite state of the art machines which provide automatic detection of explosive material, multiple views of X-ray images or liquid explosive detection systems, requirements on aviation security screeners were rather increased than reduced. According to Howell and Cooke (1989), advances in technology are associated with larger cognitive demands on human operators. While procedural and predictable tasks are processed by machines, human operators have become responsible for tasks that require inference, diagnosis, judgment, and decision making. As well in aviation security the final decision is still made by the human operator despite state of the art technology. Furthermore, nowadays that changes have to be implemented within a few days or weeks (e.g., the liquid regulation implemented in November 2006 after the terror plot in London was uncovered), the human operator is the most capable and adaptable resource in the system. Thus, the human-machine interaction and the tasks of human operators in aviation security should be investigated and adapted accordingly in order to achieve the best outcome.

In this thesis, the human factor in aviation security was investigated. Several studies that are part of the three main topics pre-employment assessment of aviation security screeners, training of aviation security screeners and age effects in aviation security screening are reported. All studies focus on the X-ray screening task which is considered to be one of the most important ones in this field. Based on the knowledge of basic object recognition theories and psychophysical methods, test procedures were developed and validated in order to select and certify the human operator for the X-ray screening task. Further studies investigate whether training can improve detection performance and whether other factors such as age or gender influence the performance additionally.

PART I: PRE-EMPLOYMENT ASSESSMENT OF AVIATION SECURITY SCREENERS. The major task of aviation security screeners is to ensure that no prohibited items are brought into the security restricted area and hence on board an airplane. To understand the single tasks and to define the basic job requirements a

job analysis was performed in PART I „TASK ANALYSIS TO DEFINE THE TASKS OF AVIATION SECURITY SCREENERS AND DEVELOP RELIABLE AND VALID PRE-EMPLOYMENT ASSESSMENT TOOLS". This analysis revealed the six main tasks X-ray screening, baggage search, body search, handling of passengers, teamwork, and coping with negative feedback. Further, a cognitive task analysis (CTA) was performed for the X-ray screening task. X-ray screening of passenger bags means to recognize prohibited items among many harmless objects that are brought along by passengers. Most object recognition models agree that recognition is defined as a process in which a stimulus representation has to match a visual memory representation. Only if this matching process leads to an activation which is high enough to exceed the internal threshold, the object is recognized (for an overview see Graf, Schwaninger, Wallraven, & Bülthoff, 2002). Thus, recognizing specific objects like threat items in X-ray images requires a visual memory representation of these objects. If a certain type of forbidden object has never been seen before, no representation in visual memory could be formed and the object is not recognizable unless it is similar to stored views of another object. Further, studies that deal with visual search, figure ground segregation and mental rotation could show that these processes influence detection performance as well (Wolfe, 1994; Wolfe, Oliva, Horowitz, Butcher, & Bompas, 2002; Palmer, 1999; Tarr & Bülthoff, 1995; Tarr, 1995). These findings from basic research studies are consistent with descriptions of experienced aviation security screeners. According to them the appearance of threat objects in X-ray images have to be learned and thus objects which are not known cannot be found. Further, detection of threat items becomes more difficult if the bag is very close-packed and many other objects distract attention. Further, objects can be superimposed by other objects in the bag or shown in an unusual view which impedes the detection as well. Both, findings from basic research studies and the CTA proposed that the X-ray screening task involves both, knowledge-based and image-based factors. Whereas the knowledge has to be learned on the job, image-based factors are rather related to visual abilities which a job applicant should already possess before getting employed. The distinction between knowledge- and image-based factors was further defined in the second study of PART I „AVIATION SECURITY SCREENERS VISUAL ABILITIES AND VISUAL KNOWLEDGE MEASUREMENT". Whether visual abilities and visual knowledge are in fact two different factors in the X-ray screening process, whereof only one is relatively independent of experience was tested in this study. Therefore, detection

performance of experienced experts and novices was measured using two X-ray screening tests. One test, the Prohibited Items Test (PIT) includes a large variety of prohibited items and measures the knowledge about the visual appearance of threat items in X-ray images of passenger bags. The other one, the X-Ray Object Recognition Test (X-Ray ORT) measures the ability to cope with image-based factors. Bag complexity, superposition and the viewpoint of threat items are defined as image-based factors. For the interpretation of X-ray images these factors always play along. It is obvious that the detection of a threat item becomes more difficult if it is in a close-packed bag, superimposed by other objects in the bag, or shown in an unusual view. In the X-Ray ORT all three image-based factors are varied systematically and the threat items are well known to everybody. Results revealed large differences between experts and novices for the PIT, but not for the X-Ray ORT. Thus, the assumption could be verified that the knowledge about which items are prohibited and what they look like in X-ray images is dependent on experience and training compared to image-based factors which are related to visual abilities. These results lead to the assumption that measuring the visual abilities of job applicants could increase detection performance later on the job. Therefore, the X-Ray ORT was evaluated in the third study of PART I "INCREASED DETECTION PERFORMANCE IN AIRPORT SECURITY SCREENING USING THE X-RAY ORT AS PRE-EMPLOYMENT ASSESSMENT TOOL". This study tested whether the X-Ray ORT is a reliable and valid pre-employment assessment tool. Therefore, different reliability and validity measures were calculated. The results show that the X-Ray ORT measures the image-based factors reliably. Further, the two validity measures, concurrent and discriminant validity could show that the X-Ray ORT measures in fact X-ray screening processes. Moreover, criterion-related validity revealed that screeners who perform well in the X-Ray ORT also have a better detection performance on the job. The detection performance on the job was measured using threat image projection (TIP) data. Another analysis shows that screeners who were employed based on their results in the X-Ray ORT perform in fact on a significantly higher level in the PIT after half a year than their colleagues who were not employed with this test. As the image-based factors bag complexity, superposition and viewpoint of threat items are related to the general visual cognition processes visual search, figure-ground segregation, and mental rotation, it can be supposed that these factors can also be measured with general visual cognition tests that are mostly part of general intelligence test batteries. Therefore, 12 subtests that are

assumed to match the X-ray screening task best were selected and applied to job applicants. Whether this cognitive test battery (CTB) can as well predict on the job performance and therefore be used to select new job applicants was investigated in the fourth study of PART I "KEEP IT SIMPLE: COGNITIVE TEST BATTERY TO SELECT JOB APPLICANTS FOR THE X-RAY SCREENING TASK IN AVIATION SECURITY". Analysis showed the relationship between ability and detection performance later on the job using structural equation modeling (SEM). As expected, factor loadings on the latent variable ability were all substantial and significant. However, results prove that the cognitive test battery can be reduced to four subtests without losing any explained variance. Additionally, an overall analysis tested the relative importance of different factors such as ability, training, age, personality traits etc. on detection performance. In all models factor loadings for ability, training and age were invariant over time and reveal the importance of these variables. Age even displays not only a direct, but also an indirect effect on performance in X-ray screening.

PART II: TRAINING OF AVIATION SECURITY SCREENERS. As discussed at the beginning and verified by the first studies, abilities and knowledge are needed to maximize the detection performance of threat items at security checkpoints. Screeners have to be familiar with as many different forbidden items as needed to represent all different types of threat items. Whether this knowledge can be gained with job experience and on the job training only or if a specific training system is needed has to be investigated. According to Schwaninger and Hofer (2004), the detection of improvised explosive devices (IEDs) in the Hold Baggage Screening (HBS) area can be improved significantly with an adaptive training system. As IEDs are normally not seen at checkpoints and vary enormously in their visual appearance, a representation of such objects in the visual memory is rather difficult to acquire. Consistent with the general model on object recognition a training system helps to store different shapes of IEDs in the visual memory and enables a better matching process afterwards. However, it can be argued that for other objects like knives, guns, and several other prohibited items that are quite often seen at checkpoints, working experience alone suffices to store representation of these objects in the visual memory and thus training effects are smaller for these items or even not observable. The first study in PART II „THE ROLE OF RECURRENT CBT FOR INCREASING AVIATION SECURITY SCREENERS' VISUAL KNOWLEDGE AND ABILITIES NEEDED IN X-RAY SCREENING" investigates whether an adaptive training system in the cabin baggage screening (CBS) area helps to improve the detection of prohibited

items for all kinds of threat objects. Therefore, experienced aviation security screeners' visual knowledge was tested before and after two years of training with the individual adaptive training system X-Ray Tutor (XRT). Further, this study investigated whether image-based factors which are stated to be rather independent of training in fact cannot be improved significantly within these two years. Results indicate that knowledge-based factors in X-ray screening can be improved substantially with an individual adaptive training compared to image-based factors that show only a small increase. The second study in PART II „INVESTIGATING TRAINING, TRANSFER, AND VIEWPOINT EFFECTS RESULTING FROM RECURRENT CBT OF X-RAY IMAGE INTERPRETATION" tests additionally whether training effects can be shown for all threat categories in similar ways. It can be expected that the training system is more effective for threat items that are normally not seen at checkpoints than for threat objects which are quite often brought along by passengers. Further, the design of this study enables to examine transfer effects, i.e. whether the training system enables a transfer of knowledge on new items that are similar in their appearance, but not trained. Last, it should investigate whether different views can be stored in visual memory and hence the viewpoint effect decreased with training. Increase in detection performance was found for the group which trained with the adaptive training system, but not for the control group which received a conventional training. For the XRT group increases could be found for all threat categories. However, this increase differed depending on the threat category. Whereas the detection performance increase was large for the category IEDs and the category Other, only small differences before and after training could be found for knives and guns. This result implies that job experience has already an effect on the detection of threat items. Threat objects like guns and knives which are relatively often taken along by passengers and therefore more often seen at checkpoints are detected better than IEDs which are normally not seen on the job. Nevertheless, the effect of job experience is quite small compared to the training effect which can be achieved with an adaptive CBT. Further, training helps to transfer the gained knowledge to other untrained objects that are similar in shape. That is, an effective training enables the screener to detect similar, but unfamiliar objects. In this study, large viewpoint effects were revealed. This is consistent with object recognition studies (Graf et al., 2002; Hayward, 2003; Tarr & Bülthoff, 1995, 1998). After training, both views were better recognized than without training. However, no significant interaction could be found. That is, contrary to the expectation, the

detection performance for difficult views was stable even after six month of training. However, it must be pointed out that the XRT training algorithm only provides the screeners with unusual views of objects once a screener can detect a prohibited item well when depicted from the easy perspective. Thus, it is unclear whether a significant interaction between viewpoint and measurement would have been observed if the training duration would have been increased (e.g., to one year).

PART III: AGE EFFECTS IN AVIATION SECURITY SCREENING. For the X-ray screening task factors such as age, gender, personality traits and working engagement can be considered to influence the detection performance. However, results indicated that gender and personality traits show rather small effects and are therefore not further reported. However, it should be considered that these factors could be rather important determinants for other tasks at the security checkpoint. In contrast the influence of age on detection performance in X-ray screening was unexpectedly high. Moreover, so far little is known about the long-term effects of intensive job-specific training of fluid intellectual abilities in cognitive aging research. Effects of job experience and training on detection performance of aviation security screeners are covered in PART III of this thesis. Study one in PART III "USE IT AND STILL LOSE IT: THE INFLUENCE OF AGE AND JOB EXPERIENCE ON DETECTION PERFORMANCE IN X-RAY SCREENING" investigates if age related declines can be found in the X-ray screening task and whether these declines can be compensated with job experience. Whereas many previous studies reported a reduced age effect on performance by means of job experience, no positive effects could be found for the X-ray screening task. Our results show clear age effects for a speeded visual search task and on the job performance which cannot be compensated with on the job experience. However, in Experiment 2 screeners who are employed for a longer time seem to profit more from experience in one of the two measurement conditions. With sufficient amount of practice they might be able to achieve similar performance level as young screeners. This raises the question whether the age effect could be reduced with training rather than experience. As the X-ray screening task requires screeners to find threat items in passenger bags that are mostly relatively seldom seen at checkpoints it could be expected that experience alone is not sufficient. Thus, the second study in PART III "TRAIN IT OR LOSE IT: THE INFLUENCE OF AGE AND TRAINING ON DETECTION PERFORMANCE IN X-RAY SCREENING" examines whether this age effect can be compensated with training on both, image-based and knowledge-based factors. Results revealed that the age effect in regard to the ability

to cope with image-based factors is similar for experienced and trained aviation security screeners. However, for knowledge-based factors younger and older screeners differ. Interestingly a covariance analysis revealed even larger age effects for older screeners. Thus, training is needed to reach expertise in the X-ray screening task, but cannot reduce the age effect. Results show that older screeners already try to compensate their decline with more training hours compared to younger ones, but are still not able to reach the same level. However, it should be noted that large individual differences were found in both studies of Part III. There are as well older screeners who perform on a significantly higher level than their younger colleagues. As the age effect seems to be an important determinant in this X-ray screening task it certainly remains to investigate whether age effects are comparably large when a screener got employed in younger days. Further analysis should test whether mandatory training hours should be adapted taking age into account.

References

Graf, M., Schwaninger, A., Wallraven, C., & Bülthoff, H. H. (2002). Psychophysical results from experiments on recognition & categorisation. *Information Society Technologies (IST) programme*, Cognitive Vision Systems – CogVis (IST-2000-29375).

Hayward, W. G. (2003). After the viewpoint debate: where next in object recognition? *Trends in Cognitive Sciences, 7(10),* 425–427.

Howell, W. C., & Cooke, N. J. (1989). Training the human information processor: A look at cognitive models. In I. L. Goldstein (Ed.), *Training and development in work organizations: Frontiers of industrial and organizational psychology* (pp. 121-182). San Francisco: Jossey-Bass.

Palmer, S. E. (1999). *Vision science – photons to phenomenology.* Cambridge, Massachusetts: the MIT Press.

Tarr, M. J. (1995). Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin and Review, 2,* 55-82.

Tarr, M. J., & Bülthoff, H. H. (1995). Is human object recognition better described by geon structural descriptions or multiple views? Comment on Biederman and Gerhardstein (1993). *Journal of Experimental Psychology: Human Perception and Performance, 21,* 1494-1505.

Tarr, M. J., & Bülthoff, H. H. (1998). Image-based object recognition in man, monkey and machine. In M. J. Tarr & H. H. Bülthoff (Eds), *Object recognition in man, monkey, and machine* (pp. 1-20). Cambridge, MA: MIT Press.

Wolfe, J. M. (1994). Visual search in continuous, naturalistic stimuli. *Vision Research, 34,* 1187-1195.

Wolfe, J. M., Oliva, A., Horowitz, T. S., Butcher, S. J., & Bompas, A. (2002). Segmentation of objects from backgrounds in visual search tasks. *Vision Research, 42,* 2985-3004.

Schwaninger, A. & Hofer, F. (2004). Evaluation of CBT for increasing threat detection performance in x-ray screening. In K. Morgan & M. J. Spector, *The internet society 2004, advances in learning, commerce and security* (pp. 147-156). Wessex: WIT Press.

PART I

PRE-EMPLOYMENT ASSESSMENT IN AVIATION SECURITY

3. SELECTION AND PRE-EMPLYOMENT ASSESSMENT OF AVIATION SECURITY SCREENERS - A TASK AND COGNITIVE TASK ANALYSIS APPROACH

3.1 ABSTRACT

This study provides an overview about the development of a reliable and valid pre-employment assessment for aviation security screeners. By means of a task analysis the job and tasks respectively of aviation security screeners were examined. Then, X-ray screening of passenger bags which is considered as one of the most important tasks of aviation security screeners was investigated using cognitive task analysis (CTA). Results from several studies are summarized within the CTA[1].

3.1 INTRODUCTION

For years terrorist attacks have presented a constant threat to civil aviation and highlighted the importance of aviation security. Furthermore, threats at new dimensions demand a reliable security check not only for hold baggage, but also for passengers and their carry-on baggage. For example the terror attack on September 11, 2001 has shown a new dimension. Former terror attacks were comparable to Lockerbie 1988 when a Pan Am Flight 103 was destroyed by a bomb in the hold baggage. As a result, immediate changes regarding hold baggage screening were introduced and automated systems for detecting improvised explosive devices (IEDs) in hold baggage were developed. However, in recent past, suicide bombers have become likely and therewith the importance of enhanced security checks of passengers and their carry-on bags has become a necessity. Although new state of the art machines, such as automatic liquid detectors, millimeter waves, X-ray machines with automatic detection of explosive materials etc. facilitate the detection of threat items, the final decision is still made by human operators (screeners). Howell and Cooke (1989) have argued that with advances in technology, cognitive demands on humans are rather increased than lowered. While procedural and predictable tasks are processed by machines, the human operators have become responsible for tasks that require inference, diagnosis, judgment, and decision making. In aviation security the human operator is a critical decision-maker and

[1] I gratefully acknowledge the help of Adrian Schwaninger in preparing the manuscript. Most of this chapter was submitted for publication.

probably the most capable and adaptable resource in the system. Above all, nowadays that changes in regulations cannot be anticipated in many cases and have to be implemented within a short period of time (e.g. liquid regulation after the terror plot in London 2006 was uncovered), the human operator is an essential component of aviation security. Nevertheless, he can also be the weakest link if not skilled or trained enough. The challenge in aviation security is to ensure maximum security while keeping the workflow at checkpoints efficient. Moreover, a reliable security check demands sufficient time and sufficient human resources which is sometimes in conflict with commercial pressure security companies are facing. In order to increase security and efficiency, all relevant factors in the security process should be evaluated by means of a job and further task analysis. Once the job requirements are defined, reliable and valid measures can be developed to increase security and efficiency.

Furthermore, analyses should investigate whether the measured factors relate to relatively stable abilities or to aptitudes, i.e. the capacity to acquire the relevant competence through training. The better the relationship between abilities, aptitudes and acquirable knowledge is defined, the better the selection criteria. Factors that cannot be trained should be addressed by a pre-employment assessment procedure to ensure that only people who have the capabilities needed to fulfill the job requirements are employed.

In the following we describe how a job and task analysis can be applied in order to define the relevant job requirements and further define selection, competency assessment and training criteria.

3.2 JOB AND TASK ANALYSIS

Job and task analytic techniques are useful tools to understand the context in which the human operator's work is taking place. According to Seamster, Redding, and Kaempf (1997) a job and task analysis should be performed primarily in order to identify the important tasks and responsibilities of a specific job. The job and task overview then provides a clear description to define the most important tasks. These should further be analyzed by means of a traditional or cognitive task analysis (CTA) depending on the task demands.

Methods to perform a job or task analysis are similar and should be chosen carefully in order that they best match the goals. According to Kirwan and Ainsworth (1992)

two data collection techniques can be distinguished, subject and observation based ones. That is, inputs from personnel who are familiar with the task can be collected using verbal protocols and questionnaires (subject based) or experts can be observed on the job (observation based). Subject based methods include amongst others the critical incident technique, questionnaires, interviews and verbal protocols. Observation based data collection techniques are activity sampling and observations. Seamster et al. (1997) suggest using observations and unstructured interviews rather as part of a preliminary task analysis. For a detailed description about the application of the above mentioned data collection techniques and other task analysis procedures see Kirwan and Ainsworth (1992), Seamster et al. (1997), Jonassen, Hannum, and Tessmer (1989). The data analyses should as well be in line with the goals and match the data collection. Results can be visualized with charts, summarized using statistical methods or presented with a verbal report.

3.2.1 Job and task analysis at the security checkpoint

In the following we report a primary job and task analysis in the field of aviation security as an example. At Zurich Airport there are three areas and workplaces respectively: the cabin baggage screening (CBS), hold baggage screening (HBS) and cargo screening. All screeners are assigned to one workplace only. Because of the separated workplaces and slightly varying tasks, the areas CBS and HBS were examined separately in order to identify the primary job tasks. For the cargo screening area no job analysis was performed. In the following, we report our findings for the CBS area only. However, it can be stated that except for one of the defined tasks, all main tasks for the HBS group differ only in their importance.

The major task of aviation security screeners is a reliable security check of passengers and passenger bags to ensure that no threat items or dangerous goods are brought into the security restricted area. Furthermore, this check should be done as efficiently and customer-friendly as possible. For the job analysis in the CBS area a subject and observation based data collection method was chosen. That is, inputs from personnel who are familiar with the task were collected and observation methods were applied by professionals. First, professionals observed aviation security screeners at their workstation. Further, unstructured interviews with screeners were conducted regarding their primary tasks they have to fulfill. Based on these inputs the security check can be reported as follows.

As can be seen in Figure 1.1, the security check of passengers and their carry-on bags is conducted by a crew that consists of four to six screeners. Each screener

Figure 1.1. Different working positions at the security checkpoint.

works at an assigned position for a defined time. Normally after 20 to 30 minutes positions are changed. Each position is linked to a specific task. Besides the body and baggage check, X-ray screening of passenger bags (see Figure 1.1, position 2) is one of the most important tasks at the checkpoint. Based on the X-ray image, the screener in front of the screen has to decide whether the carry-on bag is OK and can pass or has to be hand-searched. As it is well known, missing a threat item can have fatal consequences. However, rejecting too many harmless bags results in long waiting lines. Thus, for the screening process the knowledge about the appearance of threat items and visual abilities could be assumed to be rather important determinants to achieve a certain level of security without sacrificing efficiency. If the screener decides that a baggage have to be hand searched, another screener (see Figure 1.1, position 3) is responsible for the baggage check. For this task, language and communication skills are needed. First, screeners need to explain to passengers why they have to open the bag in order to obtain acceptance. Second, the conversation may gives a hint whether the passenger is a potential threat or not. Passengers could also try to bring prohibited items into the security restricted area and the airplane by wearing them on their body. Therefore, at minimum one female and one male screener are responsible for the body check (Figure 1.1, position 4). At Zurich Airport, the body check is done by means of a metal detector and a manual body search if required. In the next years, this manual body search might be

replaced partially by new technology (e.g. millimeter wave technology). Millimeter wave technology allows scanning people for the presence of threat objects. As clothing and other organic materials are translucent an image of the passenger can be provided which can then be interpreted. Whether this body search is done manually or using millimeter waves, specific abilities should be considered for both approaches. Another position is in front of the X-ray machine (Figure 1.1, position 1). The assigned screener has the responsibility to inform passengers about the security check and place their bags on the belt. This is done by the aviation security screener to ensure that bags are placed randomly on the belt. Otherwise a terrorist could place his bag that way that a prohibited item becomes very hard to detect in the X-ray image. In addition, this procedure ensures as well that the distance between two bags is large enough and thus enough time is given for the operator who has to interpret the X-ray image. As well for this task language and communication skills are needed. Thus, for nearly all positions in the CBS area, dealing with passengers is important. As a result, the communication between aviation security screeners and passengers can be assumed to play a key function to ensure an efficient workflow. Hence, language and communication skills, as well as customer service skills were defined as basic job requirements. Moreover, a crew always consists of several screeners who have to work as an efficient team, not only during normal operations, but especially in stressful situations. The factor teamwork becomes even more important taking into consideration that screeners are randomly assigned to a crew and especially at bigger airports do not know each other very well. Furthermore, passengers are often not pleased to pass the security control and therefore screeners have to be very patient and need the ability to cope with negative feedback even if they do their job very well. Thus, especially social skills, but also specific personality factors such as emotional stability, openness, extraversion etc. should be taken into account when people get employed.

To sum up, the job analysis revealed the six tasks X-ray screening, baggage search, body search, handling with passengers, teamwork and coping with negative feedback to be important in the CBS area.

3.3 COGNITIVE TASK ANALYSIS IN X-RAY SCREENING

Based on the job analysis, job specific knowledge and abilities should be identified that are needed to perform the single tasks proficiently in a next step. Depending on the task, a traditional (behavioral) task analysis or a CTA can be applied. Whereas

the traditional task analysis focuses mainly on procedural tasks that are noncritical, the CTA identifies and describes cognitive elements, processes such as decision making, problem solving etc. as well as knowledge and skills that are required for similar job components Seamster et al. (1997). According to Seamster et al. (1997) behavioral task analysis describes the task in terms of time spent, criticality and frequency. In contrast, CTA should be used for high performance tasks which require large amounts of knowledge or information, significant decision making or problem solving, heavy workload or time pressure, multi-tasking, substantial changing situations or considerable teamwork. Generally, these cognitive processes are more difficult to study because they are not directly observable. Moreover, traditional task analysis and CTA can also be used as complementary tools. Usually traditional task analyses are used to specify the basic job tasks and precede the CTA.

As stated by Seamster et al. (1997) quite often a CTA includes a comparison of novices and experts to find out which skills and knowledge are domain specific. These findings can help to identify selection criteria for required skills and define training sessions to acquire job specific knowledge and procedures.

CTAs can be conducted either using research or operational methods. Although research methods are often considered as too complex and time-consuming in operational settings, they can be very useful if the field of application is very broad and relatively unknown. Methods to collect and analyze data are the same as described above for the job and task analysis. For more operational methods which can be implemented within few weeks or months see Seamster et al. (1997). In a final step, a report should be written that provides a clear description of the CTA. This should include the objective of the CTA, job description, analyzed tasks, participant selection, materials and procedures used for data collection, data analyses, results and conclusion.

As an example, we performed a CTA for the X-ray screening task which is considered to be one of the most important tasks within the security check. First, a primary data collection was conducted to describe the cognitive structures and processes underlying the X-ray screening task. Based on this information data collection and analysis were performed using research methods. As the job analysis was previously reported, we focus on the description of the CTA in the following.

3.3.1 Primary data collection regarding the X-ray screening task

Besides the body and baggage search, X-ray screening is a major task in aviation security. Often, screeners have only a few seconds to decide whether an X-ray image

of a passenger bag contains a threat item or not. At security checkpoints it is of utmost importance that all threat items are detected without sacrificing efficiency. If too many bags are wrongly judged as not ok and have to be hand-searched, long waiting lines at checkpoints have to be taken into account. The X-ray screening task demands large amounts of knowledge, probably visual cognition abilities, decision making, and often multi-tasking under time constraints. Because of these factors, a CTA is reasonable. It must be noted that the conducted CTA focuses on the visual task only.

In a first step, unstructured interviews, verbal reports and observations were conducted for the primary data collection. Results of these data collection techniques as well as theories from basic research studies in object recognition let assume that the screening process involves both, knowledge-based and image-based factors.

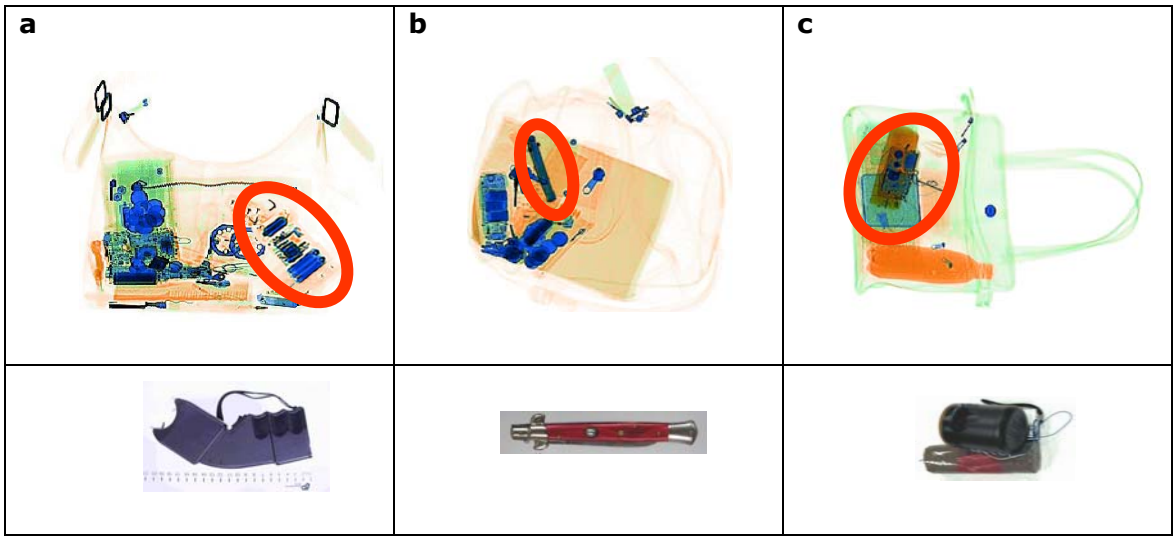

Figure 1.2. Threat items in x-ray images: (a) an electric shock device, (b) a knife and (c) an improvised explosive device (IED).

Schwaninger, Hardmeier, and Hofer (2005) defined the knowledge about which items are prohibited and what they look like in X-ray images as knowledge-based factors. X-rayed threat objects like the electric shock device depicted in Figure 1.2a can look quite different than they do in reality. Furthermore, some items in the bag look like harmless objects, but are in fact threat items that are not allowed on board an airplane (see the knife in Figure 1.2b). Again other objects like improvised explosive devices (IEDs) in Figure 1.2c are normally not seen at checkpoints and are therefore more difficult to recognize without appropriate training.

Considering the X-ray images in Figure 1.3, it can be seen that bag complexity, superposition and viewpoint of threat items in X-ray images influence the detection performance as well. The detection of a gun becomes more difficult if there are many other objects in the bag which distract attention (high bag complexity). Furthermore, it is as well more difficult if the gun is superimposed by other objects in the bag or if it is shown in an unusual view. Schwaninger et al. (2005) defined these factors as image-based factors in X-ray screening. Bag complexity, superposition and the viewpoint of threat items can as well be transferred to the visual cognition processes visual search, figure-ground segregation and mental rotation which are well known from object recognition studies. Research studies in this field revealed that it becomes harder to detect a target object if many objects are presented within a scene (e.g. Wolfe, 1994; Wolfe, Oliva, Horowitz, Butcher, & Bompas, 2002). Further, the detection is more difficult if the target object is superimposed by other objects and relevant components are not visible (Palmer, 1999). Last, viewpoint-dependent theories in object recognition predict systematic effects of viewpoint and familiarity (Tarr & Bülthoff, 1995b; Tarr, 1995). According to Palmer, Rosch, and Chase (1981) the easy viewpoint of an item refers to the canonical (i.e. easy recognizable) perspective. Thus, objects in the frontal or canonical view which is normally more often seen are easier to detect then ones presented in unfamiliar views. Although it is known from object recognition theories that different views of an object can be stored in the visual memory which facilitates the detection of rotated objects, people with good mental rotation abilities are probably more able to detect a prohibited item even if it is shown in a rotated view they have never seen before.

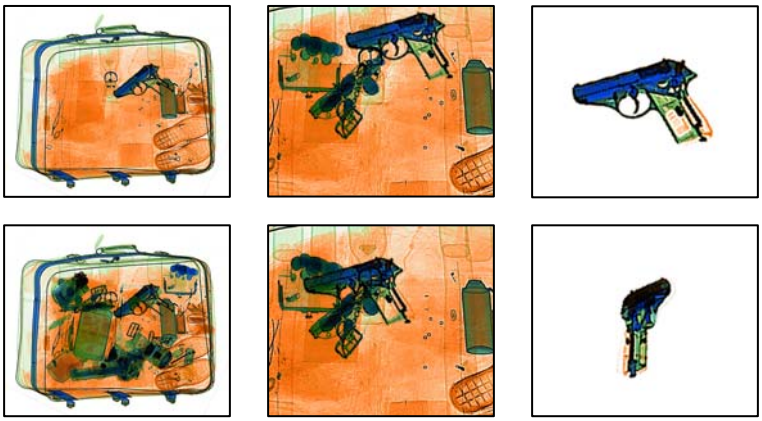

Figure 1.3. Image-based factors in x-ray images: left bag complexity, middle superposition and right viewpoint.

In state-of-the art X-ray screening equipment, different materials in X-ray images (organic, metallic materials etc.) are coded using different colors. Therefore, it can be assumed that a color vision deficiency can impair the X-ray image interpretation

task. There are different kinds of color deficiencies which are mostly inherited. The most common form is the red/green color blindness which occurs in about 8% to 12% of males and about 0.5% of females.

3.3.2 Materials and Procedure

Knowledge-based factors

An important factor in the X-ray screening task is the knowledge about which items are not allowed in the security restricted area and what they look like in X-ray images. To measure whether knowledge-based factors could in fact be improved with on the job experience or a specific training, the Prohibited Items Test (PIT) was developed. This test includes all kinds of prohibited items according to international prohibited items lists (EU, ECAC, ICAO). To keep the image-based factors constant all items are shown in the easy view only and bag complexity as well as superposition were kept relatively constant over all trials. In the PIT a total of 160 trials are shown to participants whereof 80 images are harmless bags (i.e. without any prohibited item). Each image is shown for 10 seconds on the screen. Then participants have to decide whether the bag was OK (included no prohibited item) or NOT OK (included a prohibited item) by clicking on the respective button on the screen. Further they have to indicate to which of the seven categories the prohibited item belongs to and how sure they are in their decision. For more information about this test, its reliability and validity measures see Hardmeier, Hofer, and Schwaninger (2006a).

Image-based factors

The defined image-based factors bag complexity, superposition, and viewpoint of threat items were supposed to be important determinants for the interpretation of X-ray images. As the image-based factors are related to visual search, figure-ground segregation and mental rotation, these processes can also be measured with intelligence tests which mostly include subtests about visual cognition abilities.

Another possibility is to apply an X-ray screening test which measures the defined image-based factors in X-ray images independent of knowledge. To this end, the X-Ray Object Recognition Test (X-Ray ORT) was developed. This test includes a total of 256 X-ray images of passenger bags. In this test only guns and knives, object shapes that are well known by novices are shown. Furthermore, all X-ray images are displayed in grayscale as the meaning of color in X-ray images is not known to novices. All three image-based factors in the test are varied systematically. A total of

eight different guns and eight different knives are used. All of them are shown in an easy and rotated view. Each of these items is then placed twice in a bag with high complexity level and in a bag with low complexity level whereof in one of them with high superposition by other items and the other one with low superposition. Thus, all factors are combined with each other and each threat item is shown once in each possible combination. The X-Ray ORT is a computer based test which is very easy to use. Test participants receive a short introduction which explains the test, as well as some exercise trials to familiarize with the test taking procedure. In order to ensure that object shapes are known, all guns and all knives are shown for 10 seconds on the screen before the test starts either in the frontal or rotated view. All images are displayed for 4 seconds on the screen only. Then participants have to decide whether the bag was OK (contained no gun and no knife) or NOT OK (contained a gun or a knife) by clicking on the respective button on the screen. Additionally, they have to indicate how sure they are in their decision[2] by using a slider control. The test itself is subdivided into four parts and after each part participants can take a short break if desired.

Color vision

Whether people have color deficiencies can be measured with a color blindness test. An easy to use and often applied test is the Ishihara color blindness test (Ishihara, 2005). This test consists of a series of pictures of colored spots in which Arabic digits in slightly different colors are embedded. These digits can easily be seen with normal color vision, but not with a color vision deficiency. This test is very easy to apply and takes only a few minutes to complete.

Procedure

To investigate whether knowledge-based and image-based factors can be distinguished, the detection performance between experts and novices was compared in the PIT and the X-Ray ORT. Schwaninger et al. (2005) expected larger effects between both groups for knowledge-based factors that are more related to specific knowledge than for image-based factors that are measured with the X-Ray ORT. To further test and verify these results of the first study, the detection performance of aviation security screeners in both tests was compared before and after two years of individually adaptive computer-based training (CBT). Further,

[2] For the analysis of detection performance only OK and NOT OK responses were taken into account.

correlations between the X-Ray ORT and the PIT, on the job performance and theoretical knowledge were calculated in order to validate both tests. Last, the detection performance of screeners who were employed based on the test results in the X-Ray ORT and screeners who were employed without using the X-Ray ORT was compared.

3.3.3 Participants

Results that are going to be presented are based on 453 aviation security screeners aging between 24 and 65 years (*M* = 48.94 years, *SD* = 9.09 years). 134 novices aging between 21 and 26 years (*M* = 23.24, *SD* = 1.22), and 101 job applicants in the age between 19 and 55 years (*M* = 35.25, *SD* = 9.79) who were employed using the X-Ray ORT as pre-employment assessment tool. Depending on the analysis, the sample size had to be adjusted.

3.3.4 Data analysis

Signal detection theory provides valid methods to measure the detection performance of screeners taking the hit rate and the false alarm rate into account. A hit is a correctly identified threat item, whereas sending a harmless bag to be hand searched is called false alarm. Further, a missed threat item is defined as a miss and a correct rejection refers to a correctly identified harmless bag. Figure 1.4 shows why the hit rate alone is not a valid detection performance measure. For example screener B in Figure 1.4 reaches a hit rate of 90% by simply judging most bags as NOT OK. This can be seen considering the high false alarm rate of nearly 80%. In contrast screener A reaches the same hit rate, but with a very low false alarm rate (about 10%). Thus, screener A guarantees security without sacrificing efficiency. The detection performance measures d’ and A’ take the hit and false alarm rate into account. D’ equals z(pHit) – z(pFA alarm) whereas pHit refers to the hit rate, pFA to the false alarm rate and z to the z-transformation (Green and Swets, 1966). D’ of screeners is related to the receiver operating characteristics (ROC) curves which can be seen in Figure 1.4. These curves show how the hit rate of a screener changes as a function of changes in the false alarm rate. As an example, d’ of screener A with 2.5 is remarkably higher than d’ of 0.5 of screener B.

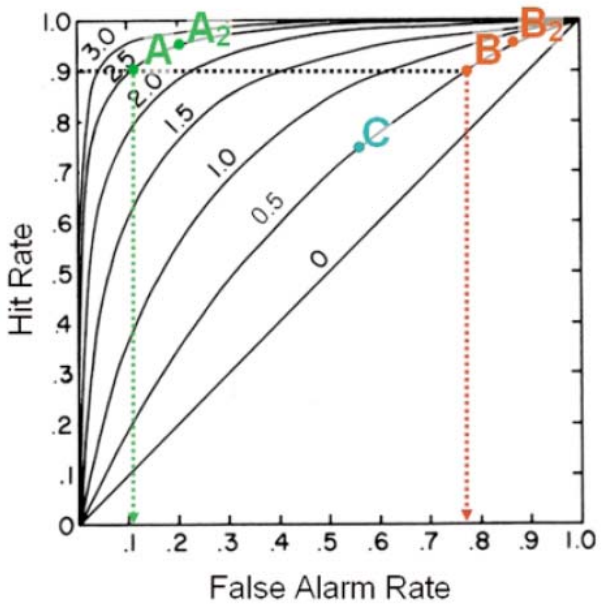

Figure 1.4. Detection performance measure d'.

Furthermore, the detection performance measure d' is independent of subjective response bias. The response bias can vary depending on the personality of screeners, cost and benefits etc. As an example if a screener judges more bags as NOT OK due to a test situation he/she changes his/her position on the ROC curves while d' remains the same (see also A2 and B2 in Figure 1.4).

In the reported studies we calculated either the detection performance measure d' or A'. In order to enable a basis of comparison, all results are calculated using d'.

3.3.5 Results

Effects of experience and training on knowledge-based and image-based factors

To investigate to what extent knowledge-based and image-based factors differ, Schwaninger et al. (2005) tested the detection performance for experienced aviation security screeners and novices in the PIT and the X-Ray ORT. Results showed that detection performance between experts and novices differed remarkably in the PIT, but only little in the X-Ray ORT. This difference becomes even more evident if the relative difference between experts and novices is computed using the following formula:

$$\frac{\text{DetectionPerformance}_{\text{experts}} - \text{DetectionPerformance}_{\text{novices}}}{\text{DetectionPerformance}_{\text{novices}}}$$

As can be seen in Figure 1.5a, percentage difference between experienced screeners and novices was 94% in the PIT and only 31% in the X-Ray ORT. Further, a detailed analysis for image-based factors revealed that the detection performance decreased significantly if threat items are shown in close-packed bags, were superimposed by other items in the bag or shown in an unusual view (see Figure 1.5b and 1.5c).

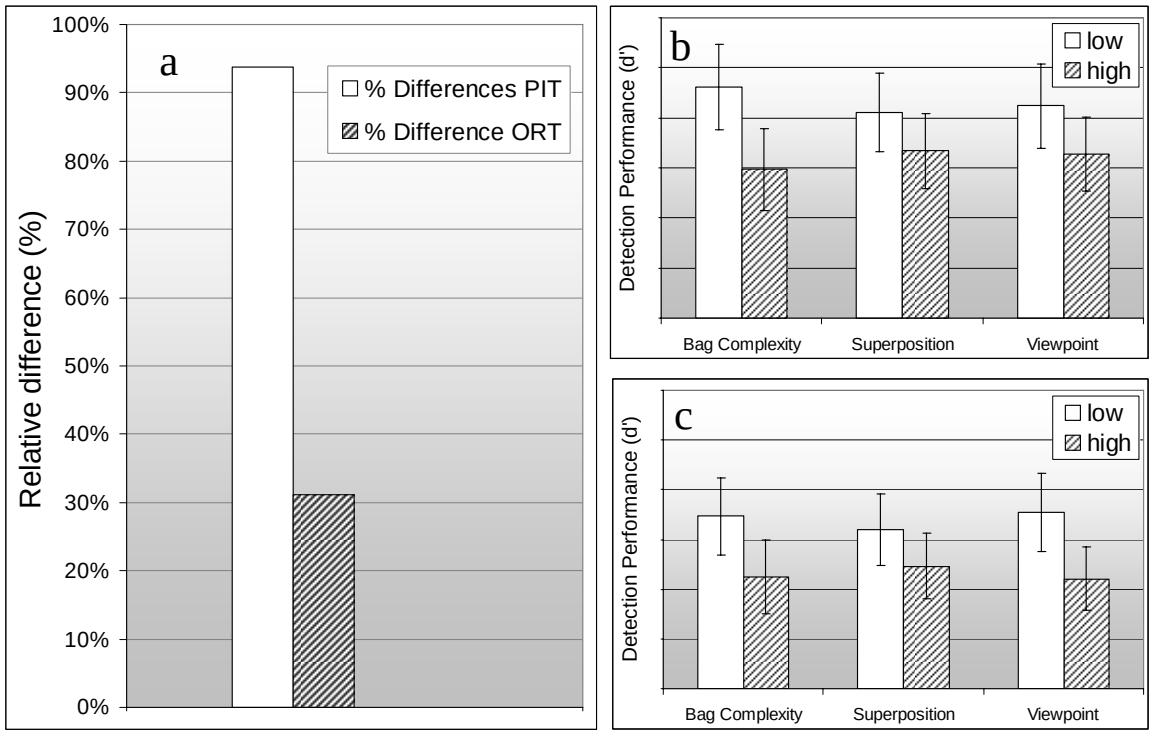

Figure 1.5. (a) Relative difference between experts and novices in the PIT and the X-Ray ORT, (b) image-based factors in the X-Ray ORT for screeners and (c) for novices. Adapted from Schwaninger et al. (2005).

Although experienced screeners perform on a slightly higher level than novices, the decreased detection performance for the three image-based factors in the difficult condition could be found for both groups. Thus, as well experienced screeners cannot compensate effects of image-based factors with their experience. Further, large individual differences could be found for both groups. Based on these test results Schwaninger et al. (2005) assumed that knowledge-based factors are indeed influenced by experience and on the job training. Further, image-based factors measured with the X-Ray ORT can rather be related to visual cognition abilities which are less dependent on experience.

To summarize, it could be shown that experience seems to influence knowledge- and image-based factors differently. Based on this study and the results of the primary data collection phase it was assumed that experience and on the job training alone is probably not enough to store the appearance of all kinds of prohibited items in the visual memory. First, some threat items are seldom or normally not seen at checkpoints (e.g. IEDs). Second, other prohibited items look quite different than in reality (e.g. electric shock device). Therefore, it could be assumed that a training system that enables to see many prohibited items within a training session helps to store all kinds of prohibited items in the visual memory. An individually adaptive training system which includes all kinds of prohibited items in different views was developed (Schwaninger, 2004, 2005b). Whether this training system helps to increase detection performance of aviation security screeners in the PIT and the X-Ray ORT after two years of individually adaptive training (2 times 20 minutes per week) was investigated by Hardmeier, Hofer, and Schwaninger (2006b). As can be seen in Figure 1.6a, the detection performance of experienced but untrained aviation security screeners (first measurement) is generally lower than the one of trained screeners (second measurement) in both tests. Further, the detection performance increase was higher for the PIT than for the X-Ray ORT. Relative difference for the PIT was 85.0%, but only 22.7% for the X-Ray ORT (see also Figure 1.6b). For more details about this study see Hardmeier et al. (2006b). Thus, an individually adaptive

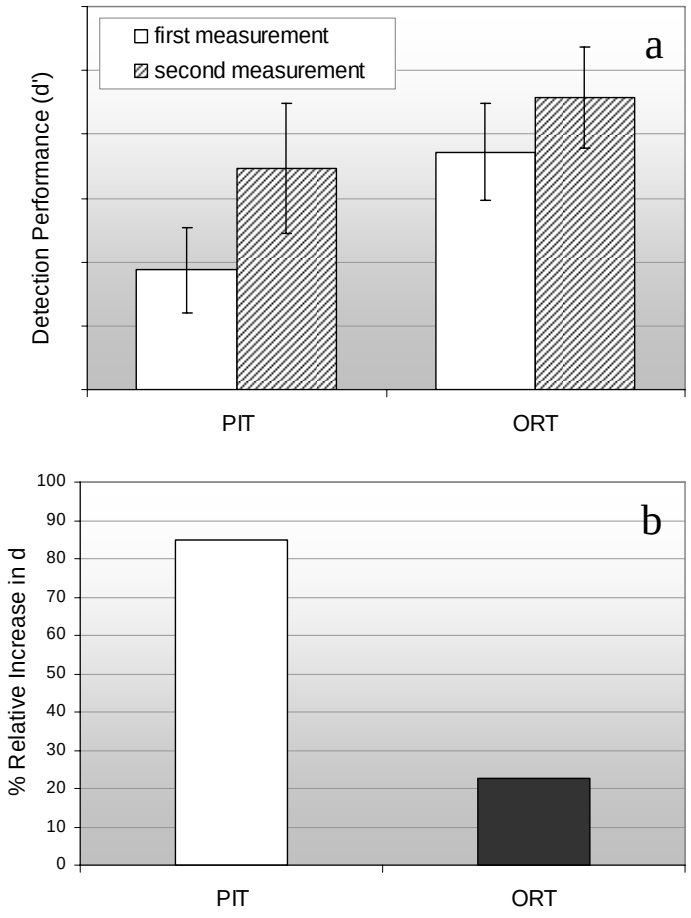

Figure 1.6. (a) Detection Performance with standard deviations in the PIT and X-Ray ORT for the first and second measurement and (b) percent difference in these two tests. Adapted from Hardmeier et al. (2006b).

CBT can strongly increase the knowledge about which items are prohibited and what they look like in X-ray images. However, it seems that image-based factors are influenced by both experience and training only to a limited extent and can therefore rather be related to individual abilities. On the basis of these results, it can be assumed that employing people with the ability to cope with image-based factors would lead to a better detection performance later on the job after training.

X-Ray ORT as pre-employment assessment tool

The rather small difference in the X-Ray ORT between novices and experienced aviation security screeners as well as between experienced and trained screeners supports the assumption that this test measures in fact visual abilities in X-ray screening. Compared to knowledge-based factors, these abilities can only be increased to some amount by experience and training. Therefore, the X-Ray ORT could be a useful instrument for pre-employment assessment purposes. Before applying this test to job applicants the X-Ray ORT was validated. Therefore, Hardmeier et al. (2006a) compared the test results in the X-Ray ORT with results in the PIT, the Computer Based Questionnaire (CBQ) and threat image projection (TIP) data. A medium correlation between both X-ray screening tests was expected as both tests deal with X-ray images. The PIT measures mainly the knowledge about prohibited items in X-ray images as image-based factors were kept relatively constant. Nevertheless, as well in the PIT image-based factors play along. However, the CBQ which is a multiple-choice test about airport specific issues and procedures at airports should be less correlated with the X-Ray ORT. Further, test results in the X-Ray ORT were correlated with TIP data. TIP is a technology that allows to measure detection performance on the job by projecting fictional threat items into real passenger bags. After each TIP image screeners receive a feedback message that a fictional threat item was present (for more information see Schwaninger & Hofer, 2004). TIP data were aggregated over a period of 17 months of 86 aviation security screeners. If the ability to cope with image-based factors is in fact important for the X-ray screening task, screeners with high ability should as well show a better detection performance on the job.

Results revealed a rather high correlation of r = .62 between the X-Ray ORT and the PIT (see Figure 1.7a). This result evidences that both X-ray screening tests measure similar processes, i.e. the interpretation of X-ray images. By contrast there was only a small correlation of r = .25 between the detection of prohibited items in X-ray images and the theoretical knowledge about airport specific issues which was measured with the CBQ, a multiple-choice test (Figure 1.7b). Regarding TIP data a medium to high correlation (r = .51) between detection performance in the X-Ray ORT and on the job performance was found (Figure 1.7c).

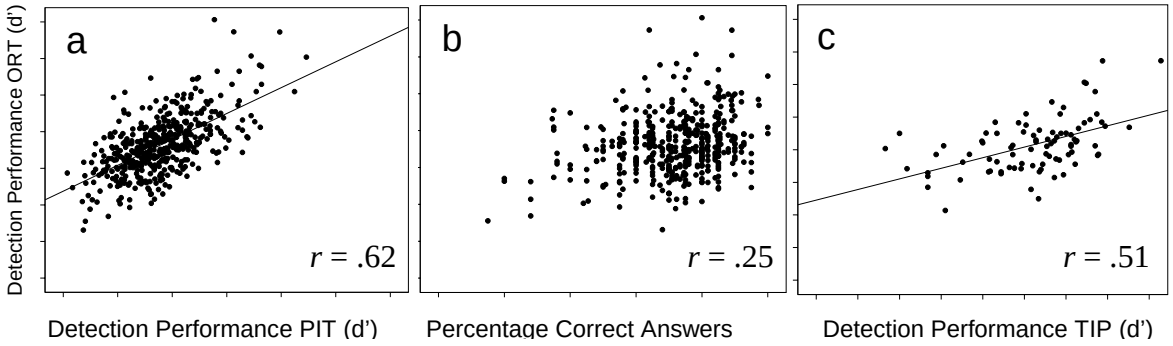

Figure 1.7. Correlation between test results in the X-Ray ORT and (a) test results in the PIT, (b) test results in the CBQ and (c) detection performance on the job (TIP).

Based on these results Hardmeier et al. (2006a) assumed that the X-Ray ORT is a valid instrument which can account for a part of the detection performance variability and therefore be used as pre-employment assessment tool to select job applicants. 101 job applicants who passed the pre-employment assessment successfully were employed as aviation security screeners. All of them had to reach a defined score in the X-Ray ORT which was clearly above the average detection performance level of novices. Further, applicants had to pass the color blindness test by Ishihara, an English and German language test, a physical exam and a job interview[3]. Whether the X-Ray ORT in fact helps to increase detection performance of aviation security screeners later on the job was also investigated by Hardmeier et al. (2006a). They compared the detection performance in the PIT of screeners who were employed with the X-Ray ORT and screeners who were employed without this test. Screeners who were not employed using the X-Ray ORT had a working experience between 2 and 26 years (M = 9.71, SD = 5.50 years). Screeners who were all hired as aviation

[3] Except for the X-Ray ORT all tests in the pre-employment assessment were as well used for the group that was selected without the X-Ray ORT.

security screeners based on the test results in the X-Ray ORT had maximum one year of working experience when taking the PIT.

The results show a significant difference between these two groups in terms of their performance measure d'. If the X-Ray ORT was used as additional selection criterion as part of the pre-employment assessment procedure, detection performance later on the job was increased significantly (see also Figure 1.8).

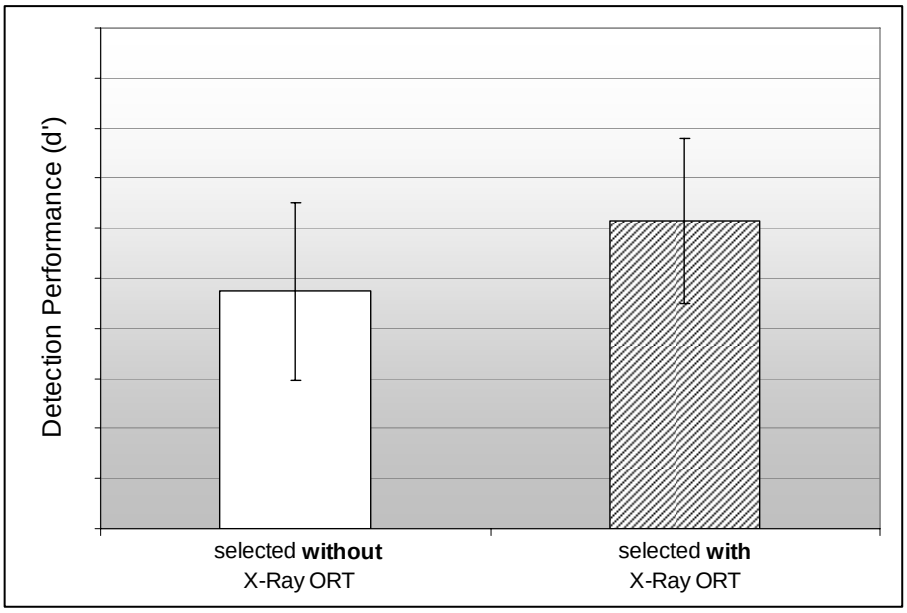

Figure 1.8. Detection performance in the PIT for screeners who were not selected using the X-Ray ORT as pre-employment assessment tool (left) selected with the the X-Ray ORT (right). Adapted from Hardmeier et al. (2006a).

3.3.6 Conclusion

Results support the assumption that image- and knowledge-based factors are important determinants in the X-ray screening task. The hypothesis that image-based factors (bag complexity, superposition, viewpoint) are related to visual abilities that are rather independent of experience and knowledge could be verified. The detection performance decreased with increasing bag complexity, superposition and unusual viewpoint of threat items for experienced and trained aviation security screeners as well as for novices. Further, large individual differences for all three groups could be found. That means, that there are large differences between people to cope with the image-based factors and these differences are still evident after individually adaptive training. Further analyses could show that results in the X-Ray ORT correlate with the detection performance in the PIT and above all with TIP data,

i.e. detection performance on the job. Therefore, the ability to cope with image-based factors should be defined as basic job requirement and be measured within a pre-employment assessment procedure. Whether the detection performance can in fact be increased using the X-Ray ORT as pre-employment assessment tool was investigated in a second step. Results evidenced, that screeners who were selected with the X-Ray ORT showed a significantly better detection performance after one year of employment compared to screeners who were not hired using the X-Ray ORT. Besides the ability to cope with image-based factors that should already be available when people get employed, the knowledge about the visual appearance of threat items is essential. In comparison with experienced aviation security screeners, novices showed a poor detection performance of prohibited items in X-ray images of passenger bags. Thus, the knowledge what threat items look like can be learned with experience and on the job training. However, a follow-up study found that an individually adaptive training system can increase the detection performance of experienced aviation security screeners enormously. As a large number of prohibited items look quite different in X-ray images than in reality, are seldom seen at checkpoints or only seen in an unusual view a training system seems to be important. Thus, an individually adaptive training system that includes a wide range of prohibited items in different views is important and should be used after employment.

To sum up, for the demanding X-ray screening task two factors should be taken into account when employing people and training employees. Only people with the relevant visual abilities should be employed. After employment screeners need to learn what prohibited objects look like in X-ray images. Therefore, an individual adaptive training which displays a large number of prohibited items within a short time should be provided.

3.4 GENERAL DISCUSSION

We could show how useful a task and cognitive task analysis respectively could be in order to understand a specific task like the X-ray screening job. However, the task analysis clearly revealed that the job of an aviation security screener includes various tasks which should be taken into account when job applicants get employed. Besides X-ray screening, baggage and body search, dealing with passengers, teamwork and coping with negative feedback were found to be important for screeners working in

the CBS area. Similar requirements could be found for the HBS area. A CTA was performed for the X-ray screening task which is supposed to be one of the most important tasks at security checkpoints.

As could be seen, X-ray screening includes different factors which should be taken into account when employing and training aviation security screeners. Whereas knowledge-based factors are part of screeners training, abilities should already be available when people get employed. One possibility to measure the ability to cope with image-based factors in X-ray screening is the X-Ray ORT. This test is a reliable and valid instrument to measure the ability to cope with bag complexity, superposition and viewpoint of threat items in X-ray images. Whether these abilities can also be measured with visual cognition tests that are part of general intelligence tests have to be examined in further research studies. These tests should include processes that reflect the X-ray screening task, thus processes such as visual search, figure-ground segregation, mental rotation, vigilance and probably logical thinking should be taken into consideration. Moreover, further research studies should investigate whether color blindness impairs the detection performance. As different materials in X-ray images, such as organic, metallic material etc. are coded using different colors, it is assumed that color vision is an important determinant for the X-ray screening job. So far, several airports conduct a color vision test before employment to ensure that the detection performance in X-ray images and thus the security is not impaired by some unknown factors. However, so far no research study have be done in order to investigate whether people with color vision deficiencies perform in fact worse in the X-ray screening task. Thus, the color vision test is a precautionary measure which is important but should be confirmed by a scientifically research study.

Further studies should clarify whether the remaining defined tasks such as baggage and body search, dealing with passengers, teamwork and coping with negative feedback are more related to abilities and have to be clarified within a pre-employment assessment or can be learned on the job.

4. AVIATION SECURITY SCREENERS VISUAL ABILITIES AND VISUAL KNOWLEDGE MEASUREMENT

4.1 ABSTRACT

A central aspect of airport security is reliable detection of forbidden objects in passenger's bags using X-ray screening equipment. Human recognition involves visual processing of the X-ray image and matching items with object representations stored in visual memory. Thus, without knowing which objects are forbidden and what they look like, prohibited items are difficult to recognize (aspect of visual knowledge). In order to measure whether a screener has acquired the necessary visual knowledge, we have applied the prohibited items test (PIT). This test contains different forbidden items according to international prohibited items lists. The items are placed in X-ray images of passenger bags so that the object shapes can be seen relatively well. Since all images can be inspected for 10 seconds, failing to recognize a threat item can be mainly attributed to a lack of visual knowledge.

The object recognition test (ORT) is more related to visual processing and encoding. Three image-based factors can be distinguished that challenge different visual processing abilities. First, depending on the rotation within a bag, an object can be more or less difficult to recognize (effect of viewpoint). Second, prohibited items can be more or less superimposed by other objects, which can impair detection performance (effect of superposition). Third, the number and type of other objects in a bag can challenge visual search and processing capacity (effect of bag complexity). The ORT has been developed to measure how well screeners cope with these image-based factors. This test contains only guns and knives, placed into bags in different views with different superposition and complexity levels. Detection performance is determined by the ability of a screener to detect threat items despite rotation, superposition and bag complexity. Since the shapes of guns and knives are usually well-known even by novices, the aspect of visual threat object knowledge is of minor importance in this test.

A total of 134 aviation security screeners and 134 novices participated in this study. Detection performance was measured using A'. The three image-based factors of the ORT were validated. The effect of view, superposition, and bag complexity were highly significant. The validity of the PIT was examined by comparing the two participant groups. Large differences were found in detection performance between screeners and novices for the PIT. This result is consistent with the assumption that

the PIT measures aspects related to visual knowledge. Although screeners were also better than novices in the ORT, the relative difference was much smaller. This result is consistent with the assumption that the ORT measures image-based factors that are related to visual processing abilities; whereas the PIT is more related to visual knowledge. For both tests, large inter-individual differences were found. Reliability was high for both participant groups and tests, indicating that they can be used for measuring performance on an individual basis. The application of the ORT and PIT for screener certification and competency assessment are discussed[4].

4.2 INTRODUCTION

The importance of aviation security has changed dramatically in the last years. As a consequence of the new threat situation large investments into technology have been made. State-of-the-art X-ray machines provide high resolution images, many image enhancement features and even automatic detection of explosive material. However, it is becoming clear since recently that the best technology is only as valuable as the humans that operate it. Indeed, reliable recognition of threat items in X-ray images of passenger bags is a demanding task. Consider the images depicted in Figure 1.2. Each of the three bags contains a threat item that could be used to severely harm people. Even though most people would probably recognize prohibited items like the electric shock device in Figure 1.2a when depicted in a photograph, this and other threat objects are relatively hard to recognize for novices because the shape features look quite different in an X-ray image than in reality. Other dangerous items (e.g., the switchblade knife in Figure 1.2b) might be missed by a novice because they look similar to harmless objects (e.g. a pen). Several other threat objects are usually not encountered in real life (e.g., improvised explosive devices, IEDs in Figure 1.2c), which stresses the importance of computer-based training in order to achieve a high detection performance within a few seconds of inspection time (Schwaninger & Hofer, 2004).

In short, the knowledge about which items are prohibited and what they look like in an X-ray image is certainly an important determinant for detection performance. The Prohibited Items Test (PIT) has been developed to measure this knowledge-based component and it therefore contains a large number of different forbidden objects according to international prohibited items lists.

[4] A similar version was published in the IEEE Aerospace and Electronic Systems, 2005. I gratefully acknowledge the help of Adrian Schwaninger and Franziska Hofer in preparing the manuscript.

As pointed out by Schwaninger (2003a) several image-based effects influence how well threat items can be recognized in X-ray images (Figure 1.3). Viewpoint can strongly affect recognition performance, which has been shown previously in many object recognition studies (for reviews see Graf, Schwaninger, Wallraven, & Bülthoff, 2002; Schwaninger, 2005a; Tarr & Bülthoff, 1995a, 1999). Since objects are often superimposed on each other in X-ray images, the degree of superposition can affect detection performance substantially. Another image-based factor is bag complexity, which is determined by the type and number of objects in a bag.

The Object Recognition Test (ORT) has been developed to measure how well screeners can cope with such image-based factors. In order to reduce effects of visual knowledge, only guns and knives are used in this test, i.e., object shapes that are usually well known also by novices.

The purpose of this study is to investigate the role of image-based and knowledge-based factors in X-ray screening using these two different tests. To what extent screeners know which items are prohibited and what they look like in passenger bags is measured by the PIT. It includes prohibited items of different categories in X-ray images of passenger bags while keeping effects of view, superposition, and bag complexity relatively constant. The objects are displayed in an easy view with a moderate degree of superposition in bags of limited complexity during 10 seconds per image. If a participant fails to detect a threat item it is therefore rather related to a lack of visual knowledge than to an attentional failure or visual processing capacity limitations. Since many different prohibited items with shapes that are often not known from everyday experience are used in the PIT, a substantial difference in detection performance between novices and screeners could be expected. The ORT measures how well someone can cope with image-based factors such as view, superposition, and bag complexity. As mentioned above, only guns and knives are used in this test, i.e., object shapes that are well known by both screeners and novices. Therefore, smaller differences between screeners and novices might be expected for the ORT compared to the PIT. However, expertise might increase visual abilities that are necessary in order to cope with image difficulty resulting from effects of viewpoint, superposition, and bag complexity. Therefore, the effect size of the interaction between image-based effects and expertise is an important measure in this study as well.

4.3 METHOD

4.3.1 Participants

A total of 268 participants took part in this study. Half were aviation security screeners, the other half were novices.

All participants were tested with the ORT and then the PIT. The screener group consisted of 67 females and 67 males at the ages of 24 and 57 years (*M* = 41.05 years, *SD* = 7.84 years). All had undergone initial classroom and on the job training and they had at least two years of work experience in airport security screening of carry-on bags.

The novices group consisted of 134 males between 21 and 26 years (*M* = 23.24 years, *SD* = 1.22 years).

4.3.2 Materials and Procedure

Prohibited Items Test (PIT)

This test contains a wide spectrum of prohibited items which can be classified into seven categories according to international prohibited items lists. The PIT version used in this study included a total of 19 guns, 27 sharp objects, 14 blunt and hunt instruments, 5 highly inflammable substances, 17 explosives, 3 chemicals, and 13 other prohibited items (e.g., buckshot, ivory). All prohibited items were depicted from an easy viewpoint and combined with a bag of medium complexity and low superposition, so their shapes could be seen relatively well and the influence of image-based factors could be minimized. X-ray images were taken from Heimann 6040i machines and displayed in color. 68 bags contained one threat item, 6 bags contained two threat items, and 6 bags contained three threat items. Each bag was shown twice resulting in a total of 160 trials. There were four blocks of 40 trials. Block order was counterbalanced across four groups of participants using a Latin Square design. Trial order was randomized within each block. Only responses to images containing one threat item were used for statistical analyses.

The PIT is fully computer-based and starts with a self-explanatory instruction, followed by a brief training session with eight examples to familiarize the participants with the procedure. Feedback is provided after each trial only in the introductory phase. Each X-ray image was displayed for a maximum of 10 seconds in the introductory and test phases. This duration is long enough to ensure that missing a threat item can be mainly attributed to a lack of visual knowledge rather than a failure of attention. For each image, participants had to decide whether the bag was

OK (no threat) or NOT OK (threat) and indicate on a slider how sure they were in their decision (confidence ratings on a 50 point scale). In addition, participants had to indicate the threat category of the prohibited item(s) by clicking the corresponding buttons on the screen (for NOT OK decisions only). Pressing the space bar displayed the next image. As the test was subdivided into four blocks, participants were allowed to take a short break after a block was completed.

Object Recognition Test (ORT)

As explained in the introduction, Schwaninger (2003a) pointed out that image-based factors such as viewpoint, superposition, and bag complexity can substantially affect detection performance in X-ray images. The ORT has been designed to measure how well people can cope with such image-based factors rather than measuring knowledge-based determinants of threat detection performance. To this end, guns and knives with the blade open are used in the ORT, i.e., object shapes that can be assumed to be known by most people. In addition, all guns and knives are shown for 10 seconds before the test starts, which further reduces the role of knowledge-based factors in this test.

In reality, a threat object can be depicted from a difficult viewpoint in a close-packed bag and be superimposed by other objects. The X-ray images used in the ORT vary systematically in image difficulty by varying the degree of view difficulty, bag complexity, and superposition, both independently and in combination. This makes it possible to investigate main effects as well as interactions between the image-based factors. All X-ray images of the ORT are in black-and-white, as color, per se, is mainly diagnostic for the material of objects in the bag, and thus, could be primarily helpful for experts.

Eight guns and eight knives with common shapes were used. Each gun and each knife was displayed in an easy view and a rotated view to measure the effect of viewpoint. In order to equalize image difficulty resulting from viewpoint changes, guns were more rotated than knives based on results of a pilot study. Each view was combined with two bags of low complexity: once with low superposition; and once with high superposition. These combinations were also generated using two closed-packed bags with a higher degree of bag complexity. In addition, each bag was presented once with and once without the threat item. Thus, there were a total of 256 trials: 2 weapons (guns, knives) * 8 (exemplars) * 2 (views) * 2 (bag complexities) * 2 (superpositions) * 2 (harmless vs. threat images). There were four

blocks of 64 trials each. The order of blocks was counterbalanced across four groups of participants using a Latin Square. Within each block the order of trials was random.

The ORT is fully computer-based. After task instructions an introductory session followed using 2 guns and 2 knives not displayed in the test phase. Feedback was provided after each trial but only in the introductory phase. Prior to the test phase, the eight guns and eight knives used at test were presented for 10 seconds, respectively. Half of the guns and knives were shown in an easy view and half were depicted in a rotated view. At test, each object was presented in the easy and the rotated view with low and high superposition and with low and high bag complexity. Each image was displayed for 4 seconds. This duration was chosen to match the demands of high passenger flow where average X-ray image inspection time at checkpoints is in the range of 3-5 seconds. For each X-ray image, participants had to decide whether the X-ray images contained one of the guns or knives shown in the introductory phase or not (NOT OK or OK response). Confidence ratings had to be provided by changing the position of a slider (90 point scale). The next trial was started by pressing the space bar. Short breaks were possible after completing one of the four blocks.

4.4 RESULTS

It is important to take the hit rate as well as the false alarm rate into account if threat and non-threat images are used in a computer-based test requiring OK and NOT OK responses. The reason is simple: A candidate could achieve a hit rate of 100% simply by judging all bags as being NOT OK. Whether a high hit rate reflects good visual detection performance, or just a lenient response bias, can only be determined if the false alarm rate is considered, too. Psychophysics provides several methods in order to derive more valid estimates based on hit and false alarm rates. A well-known measure from signal detection theory is d' (Green & Swets, 1966). It equals z(H) – z(FA) whereas H denotes the hit rate, FA the false alarm rate and z represents the transformation into z-scores (standard deviation units). An often used "non-parametric" measure is A' (Pollack & Norman, 1964). This measure represents an estimate of the area under an ROC curve that is specified by only one data point. More specifically, A' corresponds to the average area for the two linear ROC curves that maximize and minimize the hit rate. The term "non-parametric" is a bit

misleading because it only refers to the fact that the computation of A' doesn't require an priori assumption about the underlying distributions (MacMillan & Creelman, 1991; Pastore, Crawley, Berens, & Skelly, 2003). This has sometimes been regarded as an advantage over SDT measures such as d' and Δm (for a more detailed discussion of this issue. See also Hofer and Schwaninger (2004). Although only A' data are reported in this study, it should be stressed that similar results were obtained for d' data. Moreover, correlations between A' and d' were very high for both tests and screeners groups (ORT: r = .94 for screeners and r = .97 for novices, PIT: r = .95 for screeners and r = .98 for novices, all p < .001).

The results section is organized as follows: first, ANOVA results of the ORT are presented. These analyses were conducted to investigate whether detection performance of aviation security screeners and novices is affected by image-based factors. In addition, the effect of expertise on the three image-based factors measured by the ORT was examined. Second, overall detection performance in the ORT is compared to overall detection performance in the PIT[5]. More specifically, the effect of expertise on image-based factors and knowledge-based factors is analyzed, comparing detection performance of aviations security screeners with that of novices in the two tests. Finally, the results of reliability analyses are presented which were conducted to evaluate whether the ORT and PIT can be used for measuring detection performance on an individual basis.

4.4.1 ORT and Abilities to Cope with Image-Based Factors

A' scores calculated from hit and false alarm rates of the ORT were subjected to three-way analyses of variance (ANOVA) with the three within-participants factors view, bag complexity, and superposition. Results of aviation security screeners show that there were significant main effects of view (easy vs. rotated) with an effect size of η^2 = .71, $F(1, 133)$ = 318.59, MSE = 0.003, p < .001, bag complexity (low vs. high) η^2 = .83, $F(1, 133)$ = 652.96, MSE = 0.003, p < .001, and superposition (low vs. high) η^2 = .61, $F(1, 133)$ = 203.73, MSE = 0.003, p < .001. The following two-way interactions were significant: View * superposition, η^2 = .12, $F(1, 133)$ = 17.91, MSE = 0.002, p < .001, bag complexity * superposition η^2 = .12, $F(1, 133)$ = 18.22, MSE = 0.002, p < .001. Note however, that the effect sizes of these interactions are rather low when compared to the effect sizes of the main effects. All other

[5] A' scores for the PIT were calculated using the responses to images of the following categories: guns, sharp objects, hunt and blunt instruments.

interactions were not significant. In short, there were clear main effects of view, bag complexity, and superposition with very large effect sizes (see also conventions by Cohen (1988). Some interactions reached statistical significance, but the effect sizes were relatively small when compared to the effect sizes of the main effects.

Similar results could be observed for novices. Again, there were significant main effects of view (easy vs. rotated) $\eta^2 = .76$, $F(1, 133) = 428.33$, $MSE = 0.005$, $p < .001$, bag complexity (low vs. high) $\eta^2 = .72$, $F(1, 133) = 333.14$, $MSE = 0.005$, $p < .001$, and superposition (low vs. high) $\eta^2 = .63$, $F(1, 133) = 228.09$, $MSE = 0.004$, $p < .001$. All two-way interactions were significant: View * bag complexity $\eta^2 = .06$, $F(1, 133) = 9.07$, $MSE = 0.004$, $p < .01$, view * superposition $\eta^2 = .07$, $F(1, 133) = 10.43$, $MSE = 0.004$, $p < .01$, bag complexity * superposition $\eta^2 = .15$, $F(1, 133) = 23.15$, $MSE = 0.004$, $p < .001$. The three-way interaction between view, bag complexity and superposition also reached statistical significance, $\eta^2 = .03$, $F(1, 133) = 4.14$, $MSE = 0.004$, $p < .05$. As for screeners, very large effect sizes were found for main effects whereas the interactions showed much smaller effect sizes.

Figure 2.1 shows the main effects of each of the three image-based factors, averaged across the other two factors. A comparison of Figure 2.1a (aviation security screeners) and Figure 2.1b (novices) reveals that screeners were slightly better than novices while both screener groups are substantially affected by the image-based factors view, bag complexity, and superposition. In order to examine whether expertise has a differential effect on these image-based factors, a four-way analysis of variance (ANOVA) with the within-participants factors view, bag complexity, and superposition and the between-participant factor expertise was computed. There were again

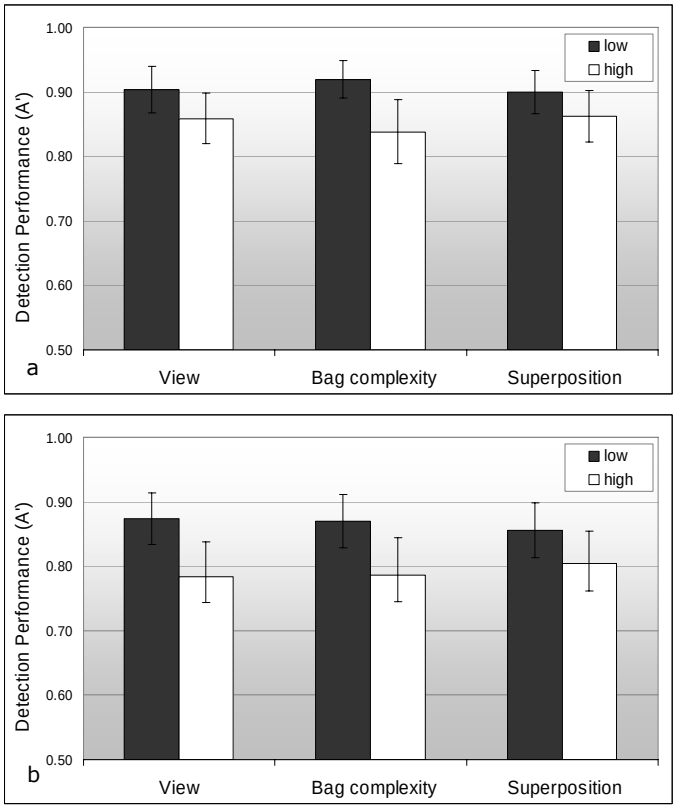

Figure 2.1. Detection performance (A') in the ORT with standard deviations (a) for aviation security screeners and (b) for novices.

significant main effects of view (easy vs. rotated) $\eta^2 = .74$, $F(1, 266) = 744.57$, $MSE = 0.004$, $p < .001$, bag complexity (low vs. high) $\eta^2 = .77$, $F(1, 266) = 884.75$, $MSE = 0.004$, $p < .001$, and superposition (low vs. high) $\eta^2 = .62$, $F(1, 266) = 428.20$, $MSE = 0.004$, $p < .001$. Two-way interactions between view and bag complexity $\eta^2 = .04$, $F(1, 266) = 10.23$, $MSE = 0.003$, $p < .01$, view and superposition $\eta^2 = .09$, $F(1, 266) = 26.17$, $MSE = 0.003$, $p < .001$, view and expertise $\eta^2 = .10$, $F(1, 266) = 30.52$, $p < .001$, and superposition and expertise $\eta^2 = .03$, $F(1, 266) = 9.39$, $p < .01$ were significant, as well as the tree-way interactions between view, bag complexity, and superposition $\eta^2 = .02$, $F(1, 266) = 5.47$, $MSE = 0.003$, $p < .05$, and bag complexity, superposition and expertise $\eta^2 = .13$, $F(1, 266) = 41.13$, $p < .001$. Although these interactions were significant, all have relatively low effect sizes when compared to the main effects. All other interactions were not significant.

In short, these results indicate that the effects of image-based factors are apparent for novices and for aviation security screeners. Further, expertise does only slightly reduce these effects of view, bag complexity, and superposition.

4.4.2 PIT, Visual Knowledge and Expertise

In contrast to the ORT, the PIT has been developed to measure whether screeners know which items are prohibited and how they look in X-ray images of passenger bags. Whereas in the ORT only guns and knives are used – object shapes that are also familiar to novices – the PIT contains all kinds of forbidden objects based on international prohibited items lists. In this test, all target objects are shown in an easy viewpoint with a moderate degree of superposition in bags of moderate bag complexity. As mentioned above, each image is shown for 10 seconds and therefore missing a threat item in the PIT can rather be attributed to a lack of visual knowledge than to an attentional failure or visual processing capacity limitations. If detection performance in the PIT is indeed mainly determined by visual experience and training with X-ray images, large differences between novices and aviation security screeners should be observed in this test. As reported in the previous section, only moderate differences between novices and screeners were found for the ORT.

In order to compare relative difference between experts and novices for the PIT and ORT, overall hit and false alarm rates were used to compute relative detection performance difference separately for the ORT and PIT using the following formula:

$(A'_{experts} - A'_{novices}) / A'_{novices}$

Relative detection performance difference between experts and novices was indeed much higher for the PIT than for the ORT (15.89% vs. 6.05%). This is consistent with the view that the PIT measures visual knowledge dependent on training and expertise, whereas the ORT measures more stable visual abilities used to cope with image-based factors such as effects of view, bag complexity, and superposition.

This main finding was further analyzed using a two-way analysis of variance (ANOVA) with the within-participant factor test type (ORT, PIT) and the between-participant factor expertise using overall A' scores from each test. There was a significant effect of test type (ORT vs. PIT) $\eta^2 = .80$, $F(1, 266) = 1075.10$, $MSE = 0.002$, $p < .001$, a significant effect of expertise (experts vs. novices) $\eta^2 = .44$, $F(1, 266) = 206.11$, $MSE = 0.004$, $p < .001$, and a significant interaction of test type and expertise $\eta^2 = .20$, $F(1, 266) = 65.30$, $p < .001$. The interaction between test type and expertise is consistent with the hypothesis that the ORT measures rather image-based factors whereas the PIT measures rather knowledge-based factors.

It must also be noted however, that correlation analyses showed that the two tests are far from being orthogonal. Overall detection performance A' of the two tests correlates with $r = .51$, $p < .001$ for experts, and $r = .42$, $p < .001$ for novices. This could at least indicate that detection performance in PIT is not only determined by visual knowledge but also by visual abilities used to cope with image-based factors as measured by the ORT.

One potential argument against the analyses of this section could be that the expert group consisted of males and females, whereas the novices group consisted only of males. However, it is unlikely that gender effects can explain the differences found between experts and novices since no significant differences were found between male and female screeners, neither for the ORT ($p = .70$) nor for the PIT ($p = .78$).

4.4.3 Reliability Analyses

Internal reliability was analyzed using Cronbach's Alpha and Guttman split-half coefficients separately for both participant groups (aviation security screeners and novices). Analyses were computed for signal plus noise trials (bags including a threat item) and noise trials (harmless bags), respectively. Reliability coefficients were computed on the basis of the percentage correct measures (i.e., hit and correct rejections), as well as on the basis of the screeners' confidence ratings (CR) for hits and correct rejections. As can be seen in Table 1.1 high reliability coefficients were found for both tests and participant groups.

Table 1.1

Reliability analyses for both tests and each group

Reliability Coefficients			PC SN	PC N	CR SN	CR N
PIT	**Screeeners**	**Cronbach Alpha**	.840	.878	.887	.924
		Split-half	.811	.915	.859	.948
	Novices	**Cronbach Alpha**	.871	.877	.885	.914
		Split-half	.882	.862	.883	.890
ORT	**Screeners**	**Cronbach Alpha**	.862	.934	.902	.962
		Split-half	.733	.813	.792	.887
	Novices	**Cronbach Alpha**	.899	.910	.916	.959
		Split-half	.778	.810	.759	.907

Note. Cronbach Alpha values and split-half reliabilities (Guttman) for both tests in each group (experts and novices separately) calculated for percentage correct (PC) and confidence ratings (CR) separately for signal plus noise (SN) and noise trials (N).

The results section has clearly shown that item difficulty in the ORT depends on the main effects and interactions between view, bag complexity, and superposition. Therefore, the high internal consistency also found for the ORT is a nice example for the fact that a test can be homogenous and multifactorial (see Kline, 2000).

4.5 DISCUSSION

The objective of this study was to examine the role of image-based and knowledge-based factors for detecting threat items in passenger bags. As pointed out by Schwaninger (2003a), image-based factors such as effects of viewpoint, bag complexity, and superposition can substantially affect detection performance. The ORT has been developed to measure how well a participant can cope with these image-based factors. This test contains guns and knives depicted in an easy and difficult view shown in bags with low and high bag complexity while being strongly or little superimposed by other objects. Main effects with large effect sizes were found for aviation security screeners as well as novices. While screeners achieved a moderately better detection performance in the ORT, they were still significantly

affected when threat items were rotated, superimposed by other objects, or shown in complex bags. This result is consistent with the view that the ORT does measure visual abilities necessary to cope with image difficulty resulting from effects of viewpoint, bag complexity and superposition. Large inter-individual differences were found both for novices as well as experts. Internal reliability was very high for both groups. Therefore, this test could be a useful tool both for competency assessment of screeners as well as for pre-employment assessment purposes.

The PIT has been developed to measure whether a screener knows which items are prohibited and what they look like in X-ray images of passenger bags. In this test, all objects are depicted in an easy view. Bag complexity and superposition are moderate so that the threat item shapes are visible. Images are shown for 10 seconds, i.e., missing a threat item can be attributed to a lack of visual knowledge rather than to an attentional failure or a visual processing capacity limitation. If the PIT is indeed related to visual knowledge based on expertise and training, large differences between novices and experts should be observed. Indeed, relative detection performance difference between novices and experts was about three times higher for the PIT than for the ORT. This result is consistent with the view that the PIT measures knowledge-based factors whereas the ORT measures visual abilities used for coping with image-based factors. As for the ORT, excellent reliability coefficients were found for the PIT. This test could therefore provide a useful tool for certification, competency, and risk assessment as well as for quality control in general.

In summary, the results of this study confirm that X-ray detection performance relies on visual abilities necessary for coping with image-based effects such as view, bag complexity, and superposition. Visual experience and training are necessary to know which items are prohibited and what they look like in X-ray images of passenger bags. Both aspects are prerequisites for a good screener and can be evaluated using the ORT and PIT.

5. INCREASED DETECTION PERFORMANCE IN AIRPORT SECURITY SCREENING USING THE X-RAY ORT AS PRE-EMPLOYMENT ASSESSMENT TOOL

5.1 ABSTRACT

Detecting prohibited items in X-ray images of passenger bags is one of the most important tasks in aviation security. This screening process includes both, knowledge-based and image-based factors. That is, the knowledge about which items are prohibited and what they look like in X-ray images (knowledge-based factors) and the ability to cope with bag complexity, superposition and rotation of the threat item (image-based factors). The X-Ray ORT was developed to measure how well screeners and novices can cope with image-based factors. Schwaninger et al. (2005) could show that image-based factors are rather independent of knowledge and therefore can only be partly enhanced through training. As these image-based factors are very important in all X-ray screening tasks, using the X-Ray ORT as pre-employment assessment tool should result in a remarkable increase in detection performance of screeners in the future. To test whether the X-Ray ORT is a useful tool to select job applicants, detection performance of screeners selected with and without the X-Ray ORT was compared in the Prohibited Items Test (PIT), which mainly measures knowledge-based factors. This means that one group of job applicants (all novices) was hired using the X-Ray ORT, whereas the other group was hired without the X-Ray ORT. Both groups of screeners had undergone initial classroom training and a minimum of one year working experience in screening carry-on baggage when they took the PIT. Results evidence that in fact detection performance in the PIT is significantly higher for the group selected with the X-Ray ORT than detection performance of screeners selected without the X-Ray ORT.
Furthermore, results reveal reliable and valid measurement of detection performance in both tests, the ORT and the PIT[6].

[6] I gratefully acknowledge the help of Adrian Schwaninger and Franziska Hofer in designing the study and preparing the manuscript for publication at the 2nd the International Conference in Air Transportation (ICRAT) 2006 in Belgrade.

5.2 INTRODUCTION

Nowadays, civil aviation has become more important and passenger flow still increases yearly. As a result, work load in aviation security increases enormously. To ensure effective and efficient work, it is very important to select and train people accurately. One of the most important tasks of aviation security screeners is detecting prohibited items such as guns, knives, improvised explosive devices (IEDs) and other prohibited items in passenger bags. During rush hours at checkpoints the decision whether a bag is OK (i.e. contains no prohibited item) or NOT OK (contains a prohibited item) has to be made within four seconds. This short time requires both, the profound knowledge about prohibited items and their appearance in X-ray images, as well as the ability to cope with image-based factors such as bag complexity, superposition and rotation of the threat item.

Referring to a general visual cognition model, recognition is defined as a successful matching of the stimulus representation with the visual memory representation. Based on this model Schwaninger et al. (2005) revealed two main factors in detecting threat items in X-ray screening, knowledge-based and image-based factors. First, screeners have to know which objects are prohibited and what they look like in X-ray images in order to recognize them (knowledge-based factors). As the appearance of prohibited items in X-ray images can differ remarkably from real life, training is very important in order to recognize them. In addition, it could be shown that an effective training system like X-Ray Tutor can significantly increase detection performance by reducing the false alarm rate. That is, through training screeners learn to distinguish reliably similar looking threat and non-threat items. Second, image-based factors influence detection performance in X-ray images enormously. Schwaninger et al. (2005) have shown three different types of image-based factors, namely bag complexity, superposition and rotation of the threat item. A threat item is more difficult to detect if it is shown in a close-packed bag as other objects can distract attention (effect of bag complexity). In addition, the more the threat item is superimposed by other objects in the bag, the harder it becomes to detect it (effect of superposition). Furthermore, a rotated threat item is more difficult to detect than a threat item shown in the frontal view (effect of viewpoint). These image-based factors are relatively independent of training and therefore rather related to visual abilities. The ability to cope with image-based factors can be measured using the X-Ray ORT. This test consists of 256 X-ray images, half of them including either a gun or a knife. These threat items are shown in the frontal and

rotated view, more or less superimposed by other objects in the bag, in a close-packed or rather empty bag.

The above described image-based factors are supposed to play a key function in the X-ray screening process. Coping with bag complexity, superposition and viewpoint of a threat item can only be partly enhanced through training and is therefore rather dependent on the ability of each screener (see also Schwaninger et al., 2005). Because these abilities play an important role in all X-ray image interpretation processes, screeners who have the relevant visual abilities should not only have a much better detection performance when untrained, but also after training and some working experience, compared to screeners who are less endued with image-based factors. To test this assumption, we compared detection performance of screeners, who were hired one year before using the X-Ray ORT with detection performance of screeners, who were selected using not fully standardized selection procedures. To compare detection performance of the two groups, we used the Prohibited Items Test (PIT). The PIT is a test including all kinds of prohibited items in X-ray images and therefore allows measuring knowledge-based factors in X-ray screening. This test provides a good possibility to measure the screener's detection performance of prohibited items independently of the selection process. Furthermore, reliability and validity of both tests, the ORT and PIT were evaluated.

5.3 METHOD

5.3.1 Participants

Two groups of aviation security screeners participated in this study. The experimental group consisted of 101 participants (71 male and 30 female) between 19 and 55 years (M = 35.25 years, SD = 9.79 years), who were all hired as security screeners based on the results of the X-Ray ORT, which was used as part of the pre-employment assessment procedure. When taking the X-Ray ORT, these job applicants had no X-ray image interpretation experience at all. Besides the X-Ray ORT, this group had to pass the color blindness test, an English test and a job interview as well in order to get employed. These screeners had about one year working experience in X-ray screening when this study was conducted (i.e. when taking the PIT).

The control group consisted of 453 screeners (141 male and 312 female) between 24 and 65 years (M = 48.94 years, SD = 9.09 years), who were hired without the X-Ray

ORT, but using an old selection procedure, which consisted of a color blindness test, an oral English test and a job interview. Working experience of these aviation security screeners varied from two years to 26 years (M = 9.71 years, SD = 5.50 years) when conducting this study (i.e. when detection performance in the PIT was compared to the experimental group).

5.3.2 Material

The X-Ray Object Recognition Test (X-Ray ORT)

The X-Ray ORT consists of 256 X-ray images and measures mainly image-based factors in X-ray screening. Therefore, only guns and knives are used as these threat objects are known by most people independent of visual experience or training and therefore are also well known by novices. Furthermore, all images are shown in black and white to eliminate color-diagnostic information for experts. To measure how good test candidates can cope with image-based factors, the image-based factors bag complexity, superposition and viewpoint are varied systematically with each other. That is, eight guns and eight knives were each combined with two bags with low complexity levels and two bags with high complexity levels, but once only little and once more superimposed by other objects in the bag. Furthermore, each bag is shown once with and once without a threat item. That is, half of the trials in the X-Ray ORT are completely harmless bags and contain neither a gun nor a knife. In the test, each image is shown for four seconds on the computer screen. Then, the test candidate has to decide whether the bag is OK (contains no gun and no knife) or NOT OK (contains a gun or a knife) by clicking the respective button on the screen. Additionally, test candidates are asked to indicate how sure they are in their decision clicking on a 50 point rating scale on the screen. For a closer description of the test design refer to Hardmeier, Hofer, & Schwaninger (2005).

Prohibited Items Test (PIT)

The Prohibited Items Test (PIT) was developed to measure how well aviation security screeners know what prohibited items look like in X-ray images. The PIT contains all kinds of prohibited items and thus measures mainly knowledge-based factors in X-ray screening. All prohibited items in the PIT can be classified into seven categories by ECAC, ICAO and EU prohibited items lists. A total of 19 guns, 27 sharp objects, 14 hunt and blunt instruments, 5 highly inflammable substances, 17 explosives, 3 chemicals and 13 other prohibited items (such as ivory, crocodile) are shown. In total the PIT includes 160 trials, half of them including prohibited items and half of

them containing no prohibited items at all. 68 of the trials containing a prohibited item included exactly one prohibited item, whereas the other twelve trials included two or three prohibited items at once[7]. As this test was developed to measure mainly knowledge-based factors in X-ray images, all threat items were shown in an easy view, combined with bags of medium complexity level and medium superposition. Thus, all three image-based factors are kept relatively constant in the PIT. Furthermore, all images were shown in color to provide a realistic test environment.

Test taking procedure in the PIT was similar to the X-Ray ORT. First, a self-explanatory instruction was shown explaining the task followed by some exercise trials to familiarize the participants with the test taking procedure. After each of the six exercise trials a visual feedback was given whether the bag was OK (contains no prohibited item) or NOT OK (contains at least one prohibited item). In the test itself no more feedback was given to the test candidates. In the PIT, all images are displayed for a maximum of ten seconds on the screen. Test candidates have to decide whether the bag contains one or more prohibited items by clicking the OK or NOT OK button on the screen. If the bag is judged as NOT OK, screeners have to indicate to which of the seven categories the prohibited item(s) belongs to by clicking on the respective button(s)[8]. Besides giving the answer OK or NOT OK, test candidates have to indicate how sure they are in their decision by clicking on a 50 point rating bar on the screen. Pressing the space bar, the next image is shown. There are four blocks of trials, after which test candidates could take an individual short break if wanted. The order of blocks is counterbalanced across four groups of participants. Within each block the order of trials is random.

5.3.3 Procedure

To test whether the X-Ray ORT is a useful pre-employment assessment tool, detection performance of screeners selected without the X-Ray ORT[9] and screeners who were hired with the X-Ray ORT was compared using the test results in the PIT. All screeners who were hired with the X-Ray ORT had completed a classroom training and about one year of working experience when taking the PIT. Experience of screeners selected without the X-Ray ORT varied between two and 26 years when

[7] This was done to assure face validity. In reality more than one prohibited item can be in a passenger bag. Note that only bags including one prohibited item were used for analysis.
[8] The answer to which of the seven categories the prohibited item(s) belonged to, was not used for the data analysis.
[9] Screeners selected without the X-Ray ORT had to take a color blindness test, as well as a common job interview.

taking the PIT (for more details on detection performance and working experience see Riegelnig & Schwaninger, 2006).

5.4 RESULTS

All test results were calculated using the "nonparametric" detection performance measure A' (see Grier, 1971; Pastore et al., 2003). A' takes into account the hit rate (i.e. bags containing a prohibited item judged as NOT OK) as well as the false alarm rate (i.e. harmless bags judged as NOT OK). This is especially important considering the task of an aviation security screener. A screener, who judges nearly all bags as NOT OK, would for sure have a high hit rate, but at the same time a very high false alarm rate and thus be very inefficient in his job. A good screener is expected to recognize most forbidden objects without being mistaken. For further information on detection performance measures, calculation and assumptions about A' see Green and Sweets (1966), Stanislaw and Todorov (1999) or MacMillan and Creelman (1991).

5.4.1 Reliability and Validity of the X-Ray ORT

Reliability of the X-Ray ORT is very high for trained aviation security screeners and novices. Cronbach Alpha values range from .887 to .966 for screeners and from .907 to .970 for novices. As well split-half reliabilities (> .781 for screeners and > .778 for novices) support reliable measurement of detection performance using the X-Ray ORT. For more details about reliability of the X-Ray ORT see Hardmeier et al. (2005). Different validity measures of the X-Ray ORT were evaluated by Hardmeier et al. (2005) in order to determine whether the test measures what it is supposed to measure and whether it can be used in making accurate decisions. Internal, convergent and discriminant validity measures evidence the former, whereas criterion-related validity refers to the correctness of decisions. Large effects of bag complexity, superposition and viewpoint could be shown for aviation security screeners and novices and support high internal validity. Furthermore, convergent and discriminant validity could be shown based on all 453 screeners selected with the old selection procedure correlating results in the X-Ray ORT with results in the PIT (r = .61, p < .001) and results in the computer-based questionnaire (CBQ) (r = .27, p < .001), respectively. The CBQ is a multiple choice questionnaire including airport specific questions about safety and security regulations at airports. Therefore, neither the ORT nor the PIT should show a high correlation with the CBQ. Criterion-

related validity was examined by correlating detection performance in the X-Ray ORT with on-the-job performance measured by Threat Image Projection (TIP) data (r = .41, p < .001). TIP systems project fictional threat images into real passenger bags during work. Therefore, TIP allows measuring on-the-job detection performance. After each TIP image screeners receive a feedback message that a fictional threat item was present. TIP data were aggregated over a period of 17 months of 86 aviation security screeners. Detection performance was calculated using A' scores, i.e. hit and false alarm rates. The correlation between the X-Ray ORT and TIP data evidences that abilities measured with the X-Ray ORT are indeed important determinants of detection performance on-the-job. For more details about calculation of these validity measures see also Hardmeier et al. (2005).

5.4.2 Reliability and Validity of the PIT

As for the X-Ray ORT, Cronbach Alpha and split-half reliabilities were calculated with 453 aviation security screeners for the PIT. All reliability measures are based on percentage corrects (PC), i.e. hits and correct rejections, as well as on confidence ratings (CR), i.e. how sure screeners were in their decision. Based on signal detection theory, reliabilities were calculated for N trials (bags without a prohibited item) and SN trials (bags with prohibited items) separately. All reliability measures are listed in Table 2.1 for the two groups of screeners separately. All values are very similar for both groups and support reliable measurement of detecting threat items in X-ray images. Cronbach Alpha values are ranging from .870 to .943 and split-half reliabilities from .864 to .944.

Cronbach Alpha values and split-half reliabilities (Guttman) of the PIT calculated for screeners selected without the X-Ray ORT (N=453) and screeners selected with the X-Ray ORT (N=101): PC = percentage correct, CR = confidence ratings, SN = signal plus noise trials, N = noise trials.

Table 2.1

Reliability analyses (PIT)

Reliability Coefficients		PC SN	PC N	CR SN	CR N
Screeners (Control Group N=453)	Cronbach Alpha	.874	.901	.910	.928
	Split-half (Guttman)	.871	.914	.900	.936
Screeners (Experimental Group N=101)	Cronbach Alpha	.908	.943	.870	.883
	Split-half (Guttman)	.878	.944	.877	.864

Note. Cronbach Alpha values and split-half reliabilities (Guttman) of the PIT calculated for screeners selected without the X-Ray ORT (N=453) and screeners selected with the X-Ray ORT (N=101): PC = percentage correct, CR = confidence ratings, SN = signal plus noise trials, N = noise trials.

Validity of the PIT can be examined calculating convergent, discriminant and criterion-related validity. These measures were calculated based on all 453 aviation security screeners who were selected without using the X-Ray ORT as pre-employment assessment tool. Convergent validity was tested correlating test scores in the PIT with test scores in the X-Ray ORT. A' scores in the PIT correlated significantly with A' scores in the X-Ray ORT (r = .61, p < .001) indicating convergent validity. This rather high correlation makes sense because both tests investigate X-ray image interpretation and obviously also in the PIT image-based factors are relevant. Furthermore, correlation between A' scores in the PIT with percentage correct answers in the computer-based questionnaire (CBQ) indicates discriminant validity (r = .26, p < .001). As for the X-Ray ORT, criterion-related validity was estimated using threat image projection (TIP) data of the same TIP-library used for the validation of the X-Ray ORT (for more details about this library please see Hardmeier et al., 2005). Correlation between test results in the PIT and on-the-job detection performance (TIP data) was r = .54 (p < .001). Thus, test results in the PIT can be used to predict on-the-job performance of screeners to a certain degree.

5.4.3 Evaluation of the X-Ray ORT as pre-employment assessment tool

In order to investigate whether the X-Ray ORT is a valuable tool for pre-employment assessment, the mean detection performance of both groups in the PIT was compared (see Figure 3.1). A significant difference in detection performance of

prohibited items between screeners selected without the X-Ray ORT and the group hired with the X-Ray ORT can be shown. The job applicants who were selected with the X-Ray ORT are significantly better in detecting prohibited items in X-ray images, $t(552) = 14.51$, $p < .001$ one year after employment. To test whether the difference in detection performance is influenced by the age of screeners or working experience (see Riegelnig & Schwaninger, 2006 for the influence of these factors on X-ray detection performance) an analysis of covariance (ANCOVA) with selection procedure as between-participants factor and age and working experience as covariates was conducted. Results show that even if these two covariates are considered, detection performance of the screeners selected with the X-Ray ORT is significantly higher compared to the other screeners, with an effect size of $\eta^2 = .07$, $F(1, 548) = 38.82$, $MSE = .004$, $p < .001$.

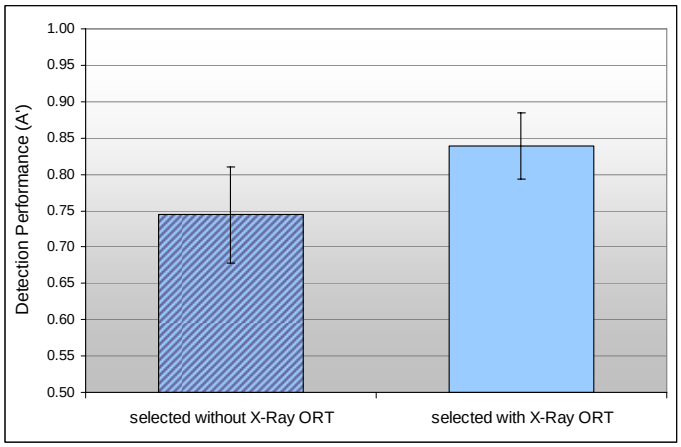

Figure 3.1. Detection performance with standard deviations in the PIT for screeners selected without the X-Ray ORT (Control Group) left and screeners selected with the X-Ray ORT (Experimental Group) right.

5.5 DISCUSSION

Overall, the results of the present study show that employing job applicants based on their results in the X-Ray ORT results in an increased detection performance in X-ray screening one year after employment, compared to detection performance of screeners selected without the X-Ray ORT. In this study, detection performance of two groups of screeners, each with a working experience of at least one year, was compared using the PIT, a computer-based X-ray screening test which measures rather knowledge-based factors in X-ray image interpretation. Compared to the screeners of the control group, who were not selected using the X-Ray ORT, job applicants who were hired based on the results in the X-Ray ORT as pre-employment assessment tool performed significantly better in the PIT one year after employment. The effect size with $\eta^2 = .07$ (Cohen, 1988) is still eminent, even when possible influences of the factors age and working experience are considered as covariates. Therefore, the ability how well someone can cope with the image-based factors, i.e.

bag complexity, superposition and viewpoint, predicts detection performance in X-ray screening to a certain degree at a later date.

In addition, statistical analyses show for both X-ray performance measurement instruments high reliability and validity. The X-Ray ORT is not only a highly reliable and valid tool for measuring how well *novices* can cope with image-based factors in X-ray images, but also how well aviation security screeners with several years of working experience can handle these factors Schwaninger et al. (2005). Furthermore, in this study we show that the PIT is a reliable instrument for measuring visual knowledge in the X-ray image interpretation task. Cronbach Alpha values were all > .87 and Guttman split half reliabilities > .86. Validity of the PIT was examined calculating convergent, discriminant and criterion-related validity. The large correlation between the PIT and the X-Ray ORT (r =.61) supports convergent validity. A rather low correlation of r = .26 between the PIT and the CBQ (a test measuring general knowledge about security issues at airports) evidence discriminant validity. Furthermore, criterion-related validity of the PIT is also quite high (r = .54).

To further investigate whether the X-Ray ORT used as pre-employment assessment tool can also predict on-the-job detection performance, Threat Image Projection (TIP) data could be measured. Currently, this is examined in a recently started study, in which the two screener groups will be compared with regard to their TIP performance. Because TIP data are only reliable when a large TIP-library with realistic images is used and data are aggregated over several months Hofer and Schwaninger (2005a), results are not yet available. As both tests, the X-Ray ORT and the PIT show high criterion-related validity, it can be assumed that the X-Ray ORT can effectively predict on-the-job detection performance.

Besides the importance of a valid and reliable pre-employment assessment procedure, intensive individual adaptive computer-based training (CBT) is also very important to improve detection performance of security screeners during work (see for example Schwaninger & Hofer, 2004 for an evaluation study of CBT). In this context, it would be interesting if screeners with high values in the X-Ray ORT show a larger training effect than screeners with low performance in the X-Ray ORT. It could be assumed that screeners who are good in coping with image-based factors profit more from training than screeners who have problems with image-based factors. This is currently also under investigation.

6. COGNITIVE TEST BATTERY TO SELECT JOB APPLICANTS FOR THE X-RAY SCREENING TASK IN AVIATION SECURITY

6.1 ABSTRACT

The job of aviation security screeners is a highly demanding task which includes specific knowledge that have to be learned on the job and visual cognition abilities. Whether these abilities can be measured within a pre-employment assessment procedure using different subtests of well established intelligence test batteries was investigated in this study. Results revealed a relationship between the latent variable ability and detection performance in X-ray screening for both samples. Moreover, a multiple group confirmatory factor analysis proved the equality of factor loadings across both samples. However, 4 of the 12 intelligence tests are sufficient to explain detection performance in X-ray screening. The influence of the X-Ray Object Recognition Test on detection performance later on the job was tested additionally. Experiment 2 investigated the strength of relationship between detection performance, age, training, ability, personality factors and working engagement among aviation security screeners. Further, direct and indirect effects were tested for the most important predictors age, training and ability[10].

6.2 INTRODUCTION

The X-ray screening task of aviation security screeners is very demanding and includes both specific knowledge and visual cognition abilities. Screeners have to acquire the knowledge about which items are prohibited and what they look like in X-ray images of passenger bags. This job and task specific knowledge and expertise respectively has to be learned after people got employed. Further, considering X-ray images different factors such as bag complexity, superposition and viewpoint of the threat items can influence the detection as well. Studies in this area could show that detection performance decreases significantly if threat items are shown in close-packed bags, if threats are more superimposed by other items and if they are shown in an unusual view. These effects were found for experts and novices. Furthermore,

[10] I gratefully acknowledge the help of Patrik Marxer and Franziska Hofer in preparing the visual cognition test battery and to Adrian Schwaninger for his help in designing the study.

large individual differences could be seen for both, experienced aviation security screeners and novices (Schwaninger et al., 2005). Schwaninger et al. (2005) defined these factors as image-based factors in X-ray screening. As they could be found for both groups, they are rather related to relatively stable abilities than training. Therefore, it can be assumed that job applicants who are able to cope with these image-based factors perform better later on the job. Thus, measuring the ability to cope with image-based factors within a pre-employment assessment should increase detection performance later on the job remarkably.

Therefore, the X-Ray Object Recognition Test (X-Ray ORT), a reliable and valid X-ray screening test that measures image-based factors relatively independent of knowledge was developed by Hardmeier et al. (2005). Results could show that test results in the X-Ray ORT correlate significantly with threat image projection (TIP) data that measures detection performance on the job. Further, aviation security screeners who were selected with the X-Ray ORT performed in another X-ray screening test that measures all kinds of prohibited items and was applied within the recurrent competency assessment significantly better than screeners who were not selected with this test (Hardmeier et al., 2006a).

However, the image-based factors should also be measurable with general visual cognition tests as these factors can be compared to the visual cognition processes visual search, figure-ground segregation and mental rotation that were investigated in many research studies. Furthermore, it can be expected that other abilities such as logical thinking, concentration or vigilance play also an important role. For example the detection of improvised explosive devices (IEDs) which vary widely in shape and form, but share a common set of components differs from the detection of other prohibited items. As not one shape as a hole has to be detected, but the three components power source, detonator and explosive material, this task probably requires rather logical thinking. Moreover, screeners have to be constantly vigilant when performing the X-ray screening task. Thus, a cognition test battery (CTB) including 12 tests that best match the X-ray screening task was applied within the pre-employment assessment additionally. Most tests are part of well established German intelligence test batteries. Four subtests of the Leistungsprüfsystem by Horn (1983), three subtests of the Intelligenz-Struktur-Test 2000 (IST 2000) by Amthauer, Brocke, Liepmann, and Beauducel (2001), the Raven's Advanced Progressive Matrices (Raven, Court, & Raven, 1980), the Frankfurter Aufmerksamkeits Inventar (FAIR) by Moosbrugger and Oehlschlägel (1996) and

three tests which were developed by the University of Zurich (Marxer, 2004) were used. Tests from the CTB were expected to measure the following unobserved latent factors figure-ground segregation, visual search, mental rotation, spatial imagination, logical thinking and vigilance.

In a first step the influence of ability on detection performance in X-ray screening was investigated using the CTB. A common factor model was estimated to measure which tests in the CTB predict on the job performance best and can therefore be used as pre-employment assessment tool. In terms of efficiency a possible shortening of the CTB was examined. Further, a full structural equation modeling (SEM) was estimated by defining the test results in the X-Ray ORT as additional indicator. We examined whether high and low performer in the X-Ray ORT can be distinguished in terms of their ability and further performance.

In a second step other factors such as training, age, personality traits, etc. which may influence detection performance were added. Finally, we tested for direct, indirect and total effects in a reduced model that includes the most important factors.

6.3 EXPERIMENT 1

In Experiment 1 confirmatory factor analyses (CFAs) were used to differentiate cognition constructs. Further, the common factor model which was estimated to measure the relationship between ability and detection performance in X-ray screening was validated by another sample and the multiple group comparison. Last, a full SEM was estimated to investigate the relationship between the general ability factor and image-based factors that were measured with the X-Ray ORT.

6.3.1 Method

6.3.1.1 Participants

The two samples used in this study consisted of 169 (M = 35.10, SD = 9.85; range 20 to 55 years) and 97 (M = 36.19, SD = 11.44; range 20 to 55 years) respectively job applicants who were employed as aviation security screeners based on their test results in the pre-employment assessment for aviation security screeners. The first sample (2006 sample) consisted of 66 females and 103 males, the second sample (2007 sample) of 51 females and 46 males. Part of the pre-employment assessment was the X-Ray ORT, the CTB, a German and English language test, a color blindness

test, a personality and work strategy questionnaire, a physical examination test and a job interview. All results except for the CTB and the two questionnaires were used as selection criteria.

6.3.1.2 Measures

Cognitive Test Battery (CTB)

The CTB consists of 12 tests which are mostly part of well established intelligence tests. All tests were conducted computer-based and not in the original paper-and-pencil form. To measure the second order factor ability, nine tests were assigned to the four first order factors figure-ground segregation, visual search, mental rotation and spatial imagination conducting CFAs. The remaining three tests Raven, Fair and a subtest of the IST 2000 (IST_MF) served as indicators.

Figure-Ground Segregation. The latent variable figure-ground segregation was measured with the subtest LPS10 of the Leistungsprüfsystem by Horn (1983) which is a major German intelligence test battery and the Noiser. The LPS10 measures the ability to recognize a shape by ignoring irrelevant other features. Participants have to choose the only simple shape out of five which fits into the complex line drawing. The test includes 40 shapes of increasing complexity. Scored is the number of correct solutions that can be answered within 3 minutes. The Noiser was developed by the University of Zurich (Marxer, 2004). It measures how well people can recognize objects that are not fully visible. The test consists of 80 line drawings of simple objects which are increasingly destroyed (level of destruction: 75%, 80%, 85% and 90%). Trials are shown for 4 seconds only and then participants have to mark the correct term out of 20 choices. Scored is the number of correct choices.

Visual Search. Visual search was measured with the Letter Search Test (LST) and the Image Comparison Test (ICT) by Marxer (2004). The LST consists of a total of 60 trials. Participants have to find a lowercase letter within three-dimensional uppercase letters. There are three difficulty levels increasing in the number of uppercase letters. Each trial is presented for 5 seconds only, then participants have to decide whether there was a lowercase letter or not. Only fifty percent of all trials contain a target object. For analysis d' is calculated. The ICT comprises of two almost identical pictures that are presented next to each other. Participants have to mark all 15 differences within 3 minutes. Scored is the number of correct marked differences.

Mental Rotation. The latent variable mental rotation was measured with the LPS7 and the Figurenauswahl (IST_FA) that are subtests of two major German intelligence test batteries, the Leistungsprüfsystem by Horn (1983) and the Intelligenz-Struktur-

Test (IST 2000) by Amthauer et al. (2001). In the LPS7 participants have to mark the flipped number or letter in a row of equal but randomly rotated numbers or letters. Participants are given 2 minutes to complete as many trials as possible out of 40. Again, scored is the number of correct solutions. The IST_FA is about rearranging several pieces to one of five possible figures. The test consists of 20 trials that have to be solved within 7 minutes. Scored is the number of correctly answered trials.

Spatial Imagination. Spatial imagination was measured with the LPS8, LPS9 of the Leistungsprüfsystem and the Würfelaufgabe (IST_WÜ) which is again a subtest of the IST 2000. The LPS8 consists of eight trials that have to be completed within 4 minutes. Participants have to mentally fold a leaf of paper into a defined form and determine for several sides which one of the leaf corresponds to the folded form. Again scored is the number of correct answers out of 40. The LPS9 measures spatial ability and asks participants to count the number of sides of three-dimensional geometric objects. Then they have to mark the correct number out of ten choices. Scored is the number of correctly marked numbers. The test duration is 3 minutes and maximum score is 40. Last, the subtest IST_WÜ consists of 20 trials that have to be completed within 9 minutes. Participants have to mentally rotate a cube and decide which of five alternatives match the target cube.

Raven. Logical thinking was measured using Raven's Advanced Progressive Matrices (Raven et al., 1980). This test measures non-verbal deductive reasoning and visual discrimination. Participants have to complete a 3 * 3 matrix of abstract figures whereof the last figure in the lower right corner is missing. They can choose the right figure out of eight alternatives. The total of 47 used matrixes increases in difficulty over time and the test duration is set to a maximum of 10 minutes. Again, scored is the number of correct solutions.

Fair. The Frankfurter Aufmerksamkeits Inventar (FAIR) by Moosbrugger and Oehlschlägel (1996) measures vigilance. The task in this test is to discriminate between very similar looking signs as fast and accurate as possible. The participants are given 6 minutes to attend the test consisting of a total of 640 trials. The number of correctly detected targets as well as correctly rejected non-targets is used for analysis.

Merkfähigkeitstest (IST_MF). The IST_MF is as well a subtest of the IST 2000 by Amthauer et al. (2001) and measures visual memory capacity. This test that measures performance of short-term memory for figures consists of 13 pairs of symbols that have to be memorized within 1 minute. Then participants have to select

the correct counterpart for all 13 symbols out of 5 alternatives within 3 minutes. Scored is the number of correct solutions.

Detection Performance in X-ray screening

The detection performance in X-ray screening was measured with two X-ray screening tests and TIP data. The prohibited items test and bomb detection test were part of the recurrent competency assessment which was conducted between 4 and 6 months after employment. Both tests are about recognizing threat items in X-ray images of passenger bags. Images were displayed for 10 and 15 seconds respectively on the screen. Then, participants have to answer whether the bag was OK (included no threat item) or NOT OK (included a threat item) by clicking on the button. Both, the prohibited items and bomb detection test differed in the 2006 and 2007 sample only insofar as other images were used. Results were calculated using d' which is a psychophysical measure and takes into account the hit and false alarm rate (Green & Swets, 1966; MacMillan & Creelman, 1991). For details about these X-ray screening tests, reliability and validity measures see Hardmeier et al. (2006a), Koller and Schwaninger (2006). TIP is a technology which allows displaying fictional threat items into real passenger bags. That way, detection performance on the job can be measured. Again, d' was calculated and used as detection performance measure. For more information about TIP data see Hofer and Schwaninger (2005b).

The X-Ray Object Recognition Test (X-Ray ORT)

The X-Ray ORT is an X-ray screening test which was developed to measure the ability to cope with image-based factors in X-ray screening relatively independent of knowledge. It consists of 256 X-ray images of passenger bags. Half of them contain either a gun or a knife. The other 128 images are harmless bags. Each bag is displayed for 4 seconds on the screen and then participants have to decide whether the bag was OK (no threat item) or NOT OK (a gun or knife) by clicking on the respective button. Detection performance was calculated using the detection performance measure d'. Test construction, its reliability and validity measures can be seen in Hardmeier et al. (2005, 2006a).

6.3.1.3 Procedure

The performance in the CTB and the X-Ray ORT was measured within the pre-employment assessment procedure. After employment all screeners had an initial training course which took three weeks. They also received training with the

individual adaptive training system X-Ray Tutor (XRT). After, screeners worked 4 to 6 months before they passed the first competency assessment which includes three X-ray screening tests and a theoretical exam on computer.

6.3.1.4 Modeling Description

The goal of this study was to test whether results in the single tests of the CTB show a relationship to detection performance in X-ray screening later on the job. The model was tested using a step-by-step procedure. First, CFAs were conducted to investigate how well the indicator variables accurately reflect the latent variables. Then, a common factor model was conducted for each group (2006 sample, 2007 sample) separately before the multiple group comparison (Byrne, 2001). Therefore, an aggregated covariance matrix was created and subjected to CFA using AMOS. As goodness-of-fit indices we report the sample-size-independent comparative fit index (CFI). Its values indicate a good fit the closer they are to one. According to Bentler (1992) values greater or equal to .90 indicate acceptable model fit. We also report the root-mean-square error of approximation (RMSEA). RMSEA values less than or equal to .05 indicate good model fit. Furthermore, the information theoretical fit measures AIC, BCC, BIC and CAIC are reported because they are less sensitive to small sample size and are not based on statistical inference using probability theory (see Arbuckle, 2005). All information theoretical fit measures should be substantially smaller than they are for the saturated model (Byrne, 2001).

6.3.2 Results and Discussion

Table 3.1 shows descriptive statistics for all indicator variables. Table 3.2 depicts the sample correlation matrix for the 2006 sample and Table 3.3 for the 2007 sample.

Table 3.1

Reliabilities, means, standard deviations of indicator variables

		2006 sample (N = 169)		2007 sample (N = 97)	
Indicator variables	Reliability	*M*	*SD*	*M*	*SD*
LPS10	.83[c] / .69[c]	0.63	0.20	0.61	0.20
Noiser	.95[a] / .91[b]	0.18	0.02	0.18	0.02
LST	.73[a] /.81[b]	0.36	0.17	0.36	0.14
ICT	.83[d]	0.65	0.18	0.64	0.16
LPS7	.83[c] / .61[c]	0.32	0.18	0.28	0.16
IST_FA	.76[a] / .79[b]	0.51	0.20	0.51	0.19
LPS8	.83[c] / .70[c]	0.63	0.31	0.62	0.29
LPS9	.83[c] / .75[c]	0.58	0.15	0.53	0.15
IST_WÜ	.80[a] / .86[b]	0.50	0.19	0.47	0.20
Raven	.93[a] /.94[b]	0.34	0.14	0.32	0.14
Fair	> .78[b] / > .85[c]	0.34	0.07	0.27	0.09
IST_MF	.92[a] / .80[b]	0.54	0.20	0.56	0.24
X-Ray ORT*	> .91[a] / > .78[b]	1.74	0.33	1.85	0.22
PIT	> .87[a] / > .87[b]	6.02	1.68	--	--
CAT	> .88[a] / > .84[b]	--	--	6.58	1.72
BDT1.0	> .80[a] / > .77[b]	3.71	2.49	--	--
BDT2.0	> .88[a] / > .80[b]	--	--	5.50	1.93
TIP	.58 - .90[b]	9.00	1.19	8.12	1.02

Note. [a] internal consistency (Cronbach alpha), [b] split-half reliability, [c] retest reliability, [d] parallel test reliability. Split-half reliability for the LPS tests was calculated for the four subtests together. Split-half reliabilities of TIP data vary depending on the image-library used. Values for the CTB are standardized and detection performance measures of all x-ray screening tests except for the X-Ray ORT have been multiplied with an arbitrary constant due to security reasons. * Reliability measures for the X-Ray ORT were based on test results from novices.

Table 3.2

Correlation matrix of indicators for 2006 sample

Indicators	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. ORT	---														
2. LST	.200	----													
3. Noiser	.106	.360	---												
4. LPS10	.150	.352	.447	---											
5. IST_MF	.155	.178	.465	.412	---										
6. Raven	.094	.357	.366	.635	.383	---									
7. Fair	.143	.159	.300	.306	.375	.361	---								
8. LPS9	.047	.325	.328	.615	.336	.598	.363	---							
9. IST_WÜ	.103	.244	.266	.331	.269	.519	.315	.401	---						
10. IST_FA	.051	.360	.263	.410	.254	.431	.175	.472	.414	---					
11. LPS8	.095	.391	.419	.587	.441	.655	.404	.640	.491	.439	---				
12. ICT	.107	.300	.300	.384	.257	.437	.203	.366	.269	.171	.452	---			
13. LPS7	.077	.212	.232	.380	.216	.429	.182	.340	.290	.326	.358	.222	---		
14. TIP	-.168	.087	.135	.187	.164	.277	.208	.258	.115	.146	.214	.131	.204	---	
15. PIT	.299	.344	.200	.189	.173	.130	.094	.186	.082	.175	.211	.122	.182	.337	---
16. BDT1.0	.345	.257	.189	.151	.145	.228	.163	.134	.183	.249	.183	.135	.206	.154	.504

Table 3.3

Correlation matrix of indicators for 2007 sample

Indicators	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. ORT	---														
2. LST	.228	---													
3. Noiser	.172	.386	---												
4. LPS10	.183	.508	.496	---											
5. IST_MF	.121	.394	.353	.396	---										
6. Raven	.147	.448	.517	.648	.440	---									
7. Fair	.115	.322	.269	.505	.441	.529	---								
8. LPS9	.244	.536	.480	.637	.440	.687	.428	---							
9. IST_WÜ	.172	.248	.289	.486	.263	.540	.307	.524	---						
10. IST_FA	.095	.392	.343	.449	.372	.442	.391	.474	.419	---					
11. LPS8	.267	.535	.495	.701	.390	.667	.452	.645	.635	.537	---				
12. ICT	-.001	.269	.359	.299	.319	.226	.214	.299	.236	.291	.390	---			
13. LPS7	.327	.259	.347	.416	.282	.399	.273	.387	.306	.270	.434	.113	---		
14. TIP	.073	-.052	.133	.297	.067	.325	.254	.259	.242	.156	.259	-.029	.345	---	
15. CAT	.192	.257	.368	.399	.159	.360	.139	.388	.242	.197	.378	.316	.383	.500	---
16. BDT2.0	.133	.288	.259	.338	.232	.395	.209	.330	.351	.220	.417	.103	.327	.480	.619

We first specified a CFA model with the four first order factors figure-ground segregation, visual search, mental rotation, spatial imagination and the three indicators Raven, Fair, IST_MF to measure the second order factor ability. However, results indicate that the second order factor loadings between the second order factor ability and the four first order factors as well as the three indicators were all not significantly different from one. Thus, all 12 indicators load on one factor and

there is no need to model separate factors. Furthermore, another first order factor named detection performance (in X-ray screening) was defined. This factor measured the detection performance in X-ray screening with the three indicators prohibited items test (PIT), bomb detection test (BDT) and TIP. As can be seen in Figure 4.1, the common factor model includes the two first order factors ability and detection performance.

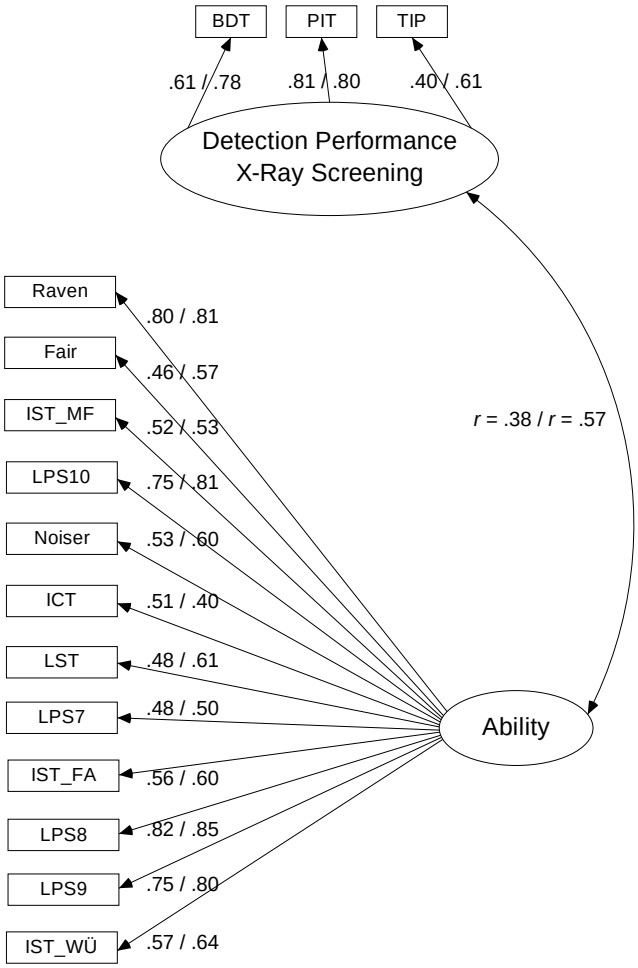

Figure 4.1. Factor model with the two factors ability and detection performance in x-ray screening (circles) and the 15 indicators. For clarity, measurement errors are omitted. Standardized loadings are indicated for the 2006 (left) and the 2007 data (right).

We first estimated the model for each group separately before conducting multiple group comparison. For the measurement model with the 2006 sample all factor loadings on the two constructs ability and detection performance were substantial

and significant. The covariance between ability and detection performance was 0.018 (*SE* = 0.006), $p < .01$, corresponding to a correlation of $r = .38$. According to Hu and Bentler (1999), Marsh, Hau, & Wen (2004) the model fit was very good and should not be rejected (see Table 3.4). As indicated by the goodness-of-fit indices, the model for the 2007 sample reproduced the covariance matrix as well very well (see Table 3.4). Covariance between the two constructs ability and detection performance was 0.027 (*SE* = 0.007), $p < .01$ and the correlation significant ($r = .57$) respectively. In both models (2006 and 2007) no substantial modifications were required.

Table 3.4

Fit Indices for Model 1 and Model 2

Model	X^2 (*df*)	CFI	RMSEA	Pclose	AIC	BCC	BIC	CAIC
Model 1 (2006 sample)	125.72(89)	.952	.050	0.497	187.72 (240.00)	194.24 (265.26)	284.74 (615.59)	315.74 (735.59)
Model 2 (2007 sample)	98.99(89)	.981	.036	0.720	160.99 (240.00)	174.58 (292.60)	238.48 (539.98)	269.48 (659.98)

Note. Saturated model values for the information theoretical fit measures AIC, BCC, BIC and CAIC are reported in brackets.

In order to test the equality of factor loadings across both samples, a multiple group confirmatory factor analysis (MGCFA) was conducted. First, we tested whether the model configuration across the two groups is equal. According Bollen (1989) configural invariance means that the same causal structure holds in several populations. The configural invariance model fitted the data very well and evidenced that all items load in fact on the same construct $X^2(178, N = 169{,}97) = 224.81$, $p < .05$, CFI = .964, RMSEA = .032, Pclose = .994, AIC = 348.81 (saturated model 480.00), BCC = 369.06 (saturated model 558.39). Second, metric invariance was tested. Steenkamp and Baumgartner (1998) defined metric invariance as precondition to compare relationships over different groups. This metric invariant model fitted even better and indicated that the indicators and their corresponding latent variables are invariant ($X^2(191, N = 169{,}97) = 232.02$, $p < .05$, CFI = .968, RMSEA = .029, Pclose = .999, AIC = 330.02 (saturated model 480.00), BCC = 346.03 (saturated Model 558.39)). Thus, factor loadings are the same. Further, the X^2 difference test showed no significant difference between the two models ($p = .89$). Moreover, the relation between ability and detection performance was equal (p

= .19) indicating no difference across the two groups. In a third step, scalar invariance was tested by constraining the intercepts of the different indicators across the groups to be equal. This is a precondition for mean comparison of the latent variables across the groups[11]. Constraining all intercepts to be equal across the groups was first not supported by the data. The intercept of the indicator TIP was significantly different across the groups implying that the lowest level of TIP was significantly higher in the 2007 group. Therefore, we released this constraint for the final model which improved the model fit considerably $X^2(203, N = 169, 97) = 1.55$, $p < .01$, CFI = .915, RMSEA = .046, Pclose = .736, AIC = 447.67 (saturated model 540.00), BCC = 469.55 (saturated model 628.19). The partially scalar invariant model still allows us to compare the means across the groups meaningfully (Byrne, Shavelson, & Muthen, 1989; Steenkamp & Baumgartner, 1998; Vandenberg 2000; 2002). The mean level of ability was not significantly different across the groups (p = .11). However, detection performance in X-ray screening was significantly higher (p < .01) for the 2007 sample which can be explained with the higher amount of individually adaptive computer based training in the 2007 sample.

In order to test what part of the variability in detection performance can be accounted for by the theoretical variables, we performed a full structural equation model analysis. Our main explanatory latent variable is ability of screeners. In addition, the test result in the X-Ray ORT is expected to account for a part of the detection performance variability. Again, a multiple group SEM was conducted to compare the effect sizes across groups. The model fit indicated with $X^2(206, N = 169, 97) = 1.33$, $p < .01$, CFI = .949, RMSEA = .036, Pclose = .986, AIC = 406.52 (saturated model 544.00), BCC = 429.68 (saturated model 639.45) a good fit. In both groups, ability and the X-Ray ORT display a significant effect on detection performance. Furthermore, these effects are not significantly different across the two groups (p = .19). Whether the effect of ability varies across individuals with high and low values in the X-Ray ORT was tested using interactions. Therefore, each sample was divided into high and low X-Ray ORT performers using a median split (Yang-Wallentin, Schmidt, Davidov, & Bamberg, 2004). Because of the small sample size, the number of indicators to measure the first order factor ability was reduced to the four indicators Raven, LPS8, LPS9 and LPS10[12]. Moreover, we used 10 percent as the

[11] In this model the mean information is used in addition to the variance-covariance information from the data matrix.

[12] There was no difference in the substantive results, especially in the prediction of the detection performance by reducing the number of indicators of ability from 12 to 4 indicators.

critical significance level. Results evidenced that the effect of ability was equal across the two groups in 2006 and 2007 for both, the low and high X-Ray ORT performers. However, the effect in the group which performed on a high level in the X-Ray ORT was significantly larger ($p < .10$) for both samples implying that there is a positive relationship between generalized ability and the ability to cope with image-based factors in X-ray screening.

6.4 EXPERIMENT 2

In Experiment 2 we tested what part of the variability in detection performance can be accounted by the six factors ability, X-Ray ORT, training, age, personality and working strategy, Furthermore, direct, indirect and total effects were tested for a reduced model which included the most important indicators.

6.4.1 Method

6.4.1.1 Participants

The main sample in this study included again the 97 (M = 36.19, SD = 11.44; range 20 to 55 years) job applicants who were employed as aviation security screeners based on their test results in the pre-employment assessment for aviation security screeners. Data from 130 aviation security screeners between 23 and 59 years (M = 45.72, SD = 8.27) who were employed for at least 4 years were used additionally (screener group).

6.4.1.2 Measures

Cognitive Test Battery (CTB)

For this analysis the cognitive test battery was reduced to the four most important indicators Raven, LPS8, LPS9 and LPS10.

Training

Training hours with an individual adaptive computer based training system were measured from employment to the first competency assessment for the 2007 sample. For the screener group overall training hours were calculated. The XRT is a training system that includes a large image library of X-ray images of passenger bags and threat items in many different views. Threat items are projected into bags based on the individually adaptive training algorithm of XRT. The XRT starts with easy views at the beginning and increases image difficulty depending on the

screeners' performance. Image difficulty is defined by more difficult viewpoints, increased bag complexity and superposition. Thereby a very efficient training can be provided to users. For more details about the training system and its evaluation see Schwaninger (2004), Koller, Hardmeier, Michel, and Schwaninger (2008).

Personality Questionnaire (NEO-FFI)

Personality was measured using the NEO-FFI by Costa and McCrae (Borkenau & Ostendorf, 1993), also called the Big Five. This test measures the five personality traits openness, conscientiousness, extraversion, agreeableness and neuroticism. In short, openness includes having wide interests, and being imaginative and insightful. People high in conscientiousness tend to be organized, thorough, and planful. Extraversion encompasses specific traits as talkative, energetic, and assertive. Agreeableness includes traits like sympathetic, kind, and affectionate. Last neuroticism is characterized by traits like tense, moody, and anxious. Each trait is measured with 15 questions. For each question, participants had to indicate if they strongly disagree, disagree, neither disagree nor agree, agree or strongly agree.

Working strategies and working experience (AVEM)

The AVEM (Arbeitsbezogenes Verhaltens- und Erlebensmuster) by Schaarschmidt and Fischer (1996) measures working strategies and working experience using 66 questions. All questions measure 11 dimensions which can be summarized to the three main factors working engagement, individual hardiness and behavior under pressure, as well as attitude towards life and healthiness. Again, questions had to be answered with strongly disagree, disagree, partly disagree and partly agree, agree, strongly agree.

6.4.1.3 Procedure

For the 2007 sample the performance in the CTB and the X-Ray ORT was measured within the pre-employment assessment procedure. Both questionnaires, the NEO-FFI and the AVEM were as well part of the pre-employment assessment. After employment all screeners had an initial training course which took three weeks. They also received training with the individual adaptive training system. On average each screeners should train 10 hours during the initial training course and thereafter at least two times 20 minutes per week. Training hours until the first competency assessment was on average 18.37 hours (*SD* = 11.60). After the initial training course, screeners worked 4 to 6 months before they passed the first competency

assessment which includes three X-ray screening tests and a theoretical exam on computer.

For the screener group the CTB was taken 4 to 6 years before the recurrent certification tests used in this study.

6.4.1.4 Modeling Description

The aim of Experiment 2 was to test the influence of other important factors on the X-ray screening task additionally. Again, the model was tested using a step-by-step procedure. First, CFAs were conducted to investigate how well the indicator variables accurately reflect the latent variables personality and working strategy and working experience. Then, a full SEM was conducted for the 2007 sample. Direct and indirect effects were then tested with both groups.

6.4.2 Results and Discussion

Descriptive statistics for all indicator variables are presented in Table 3.5. Further, the sample correlation matrix for the 2007 sample for all indicators is shown in Table 3.6.

Table 3.5

Reliabilities, means, standard deviations of indicator variables

		2007 sample (N = 97)		Screener Group (N = 130)	
Indicator variables	Reliability	*M*	*SD*	*M*	*SD*
Raven	.93[a] /.94[b]	0.32	0.14	0.27	0.14
LPS8	.83[c] / .70[c]	0.62	0.29	0.47	0.25
LPS9	.83[c] / .75[c]	0.53	0.15	0.44	0.17
LPS10	.83[c] / .69[c]	0.61	0.20	0.47	0.19
X-Ray ORT*	> .91[a] / > .78[b]	1.85	0.22	--	--
NEO-FFI_N	.85[a]	1.11	0.48	--	--
NEO-FFI_E	.80 [a]	2.83	0.41	--	--
NEO-FFI_A	.71 [a]	2.99	0.34	--	--
NEO-FFI_C	.85 [a]	3.23	0.45	--	--
AVEM_WE	> .78[a] / > .76[b]	21.14	4.49	--	--

Indicator variables	Reliability	2007 sample (N = 97)		Screener Group (N = 130)	
		M	*SD*	*M*	*SD*
Training	--	18.37	11.60	167.04	68.91
Age	--	36.19	11.44	59.62	23.85
PIT (Screener Group)	> .87[a] / > .87[b]	--	--	6.55	1.96
CAT (2007 sample)	> .88[a] / > .84[b]	6.58	1.72	--	--
BDT 1.0 (Screener Group)	> .80[a] / > .77[b]	--	--	8.40	4.27
BDT2.0 (2007 sample)	> .88[a] / > .80[b]	5.50	1.93	--	--
TIP	.58 - .90[b]	8.12	1.02	--	--

Note. [a] internal consistency (Cronbach alpha), [b] split-half reliability, [c] retest reliability, [d] parallel test reliability. Split-half reliability for the LPS tests was calculated for the four subtests together. Split-half reliabilities for TIP data vary depending on the image-library used. Values for the CTB are standardized and detection performance measures of all x-ray screening tests except for the X-Ray ORT have been multiplied with an arbitrary constant due to security reasons. *Reliability measures for the X-Ray ORT were based on test results from novices.

Table 3.6

Correlation matrix of indicators for 2007 sample

Indicators		1	2	3	4	5	6	7	8	9	10	11	12	13
1.	Age	--												
2.	Training	.24	--											
3.	AVEM	-.42	-.07	--										
4.	TIP	-.38	.10	.09	--									
5.	NEO_N	.03	.14	-.23	-.00	--								
6.	NEO_E	-.19	-.16	.27	.07	-.43	--							
7.	NEO_A	-.02	-.12	-.14	.05	-.33	.38	--						
8.	NEO_C	-.03	-.16	.27	.01	-.43	.50	.38	--					
9.	Raven	-.60	-.23	.17	.27	-.01	.01	.01	.01	--				
10.	LPS8	-.63	-.24	.18	.28	-.01	.01	.01	.01	.68	--			
11.	LPS9	-.58	-.22	.16	.26	-.01	.01	.00	.01	.63	.66	--		
12.	LPS10	-.62	-.24	.17	.27	-.01	.01	.01	.01	.66	.70	.64	--	
13.	PIT	-.58	.16	.14	.49	-.01	.10	.07	.01	.40	.42	.39	.41	--
14.	BDT	-.51	.14	.12	.43	-.01	.09	.06	.01	.35	.37	.34	.36	.65

First, CFAs of the NEO-FFI and AVEM were conducted. For the NEO-FFI a CFA showed no significant loading of the component "Openness" on the latent factor personality. Therefore, the first order factor personality was described by the four indicators

conscientiousness, extraversion, agreeableness and neuroticism. However, further analysis evidenced that the factor openness could be integrated in the full SEM as independent factor. However, we restricted to the four personality traits to avoid a too complex model with too many free parameters to be estimated compared to the sample size available. Furthermore, we calculated a CFA for the AVEM. As defined by the authors (Schaarschmidt & Fischer, 1996), we specified a model in which three first order factors were explained by the 11 indicators. We have chosen working engagement (AVEM_WE) as the most valid indicator to measure working strategy. Again, we decided to use only one indicator because of the small sample size available. As well ability was measured with four indicators only. Since the X-Ray ORT in the model was consistently not significant we have omitted this variable.

According to all fit criteria the model (see Figure 4.2) is very satisfactory X^2(63, N= 97) = 67.00, p = .34, CFI = .992, RMSEA = .027, Pclose = .764, AIC = 151.00 (saturated Model 210.00), BCC = 168.03 (saturated model 252.57), BIC = 256.00 (saturated model 472.48), CAIC = 298.00 (saturated model 577.48)[13]. All coefficients are significant on the .05 level except for the effect of ability and personality which are only significant on the 9 and 13.5 percent level respectively. However, taking into consideration that the sample size is very small and that we have no random sample from a defined population, these significant levels can be accepted as relevant. Moreover, additional analysis using bays estimation instead of maximum likelihood which is better suited for non-normal data and small sample size according to Arbuckle (2005) resulted in highly significant effects[14]. Results show that factor loadings for ability, age and training were invariant over time. The variance of ability on detection performance in X-ray screening was estimated to be .24. Thus, ability has at least a weak and positive effect on the probability of detecting prohibited items in X-ray images of passenger bags holding all other effects constant. Further, standardized regression coefficient of .44 indicate that the training effect is much higher implying the importance of training and the knowledge what prohibited items look like in X-ray images. However, the strongest effect was found for age with -.69. Thus, the older the aviation security screeners, the worse the detection performance in X-ray screening. Moreover, the effect of personality is positive, but rather weak with .16 implying that persons who are more extrovert,

[13] For the model (see Figure 2) we introduced an error correlation between the error of the indicator agreeableness and work engagement because of the high modification index which indicated a strong misfit of the model when we set the correlation to zero.

[14] Reestimation of the model with bayes procedure showed similar coefficients which were all highly significant.

conscientious, agreeable and less neurotic show increased detection performance. Results also show a slightly negative effect of work engagement on detection performance of -.21 although the bivariate relationship between detection performance and work engagement is positive. This implies a suppressor effect of the other independent variables leading to the negative coefficient between work engagement and detection performance.

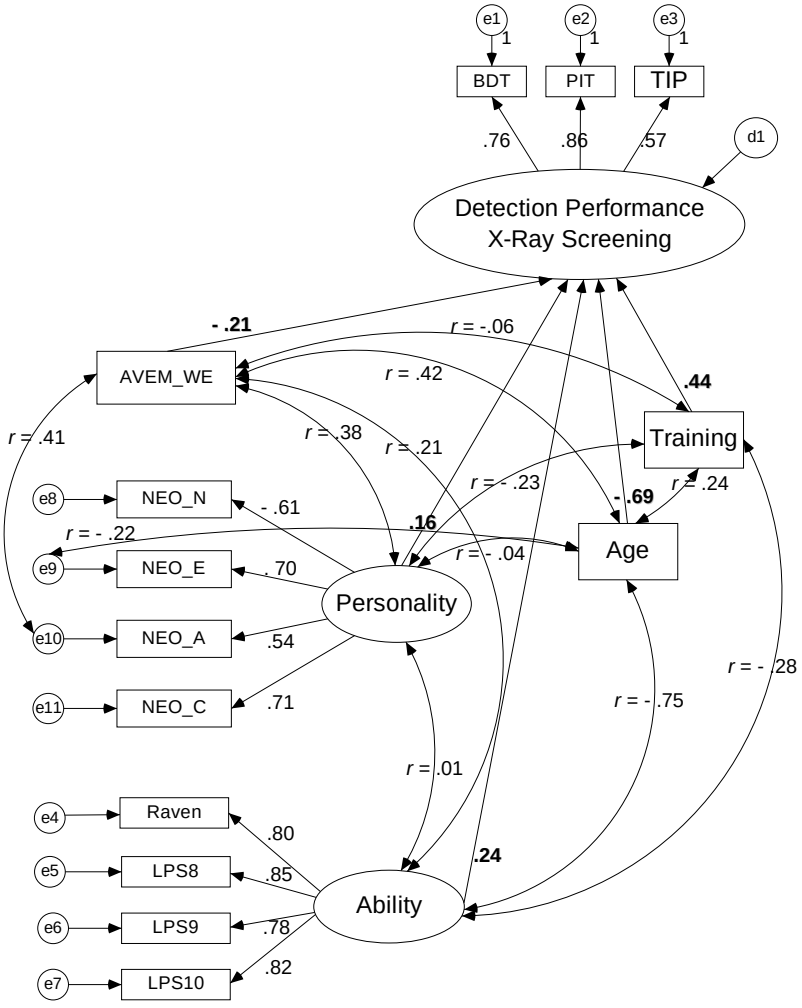

Figure 4.2. Structural model based on the 2007 data. Circles enclose factors and boxes enclose indicators. Standardized factor loadings are written in bold type.

Last, we tested for direct, indirect and total effects with a reduced model which included the three most important predictors of detection performance, namely age, training and ability. We expect a direct effect of age on detection performance and ability, as well as a direct effect of ability and training on detection performance. In

contrast, age is expected to have a negative effect on both ability and performance. This model (Figure 4.3) had a very good fit to the data $X^2(24, N= 97) = 24.32$, $p = .44$, CFI = .999, RMSEA = .012, Pclose = .707, AIC = 66.32 (saturated Model 90.00), BCC = 71.64 (saturated Model 101.39), BIC = 118.82 (saturated model 202.49), CAIC = 139.82 (saturated model 247.49).

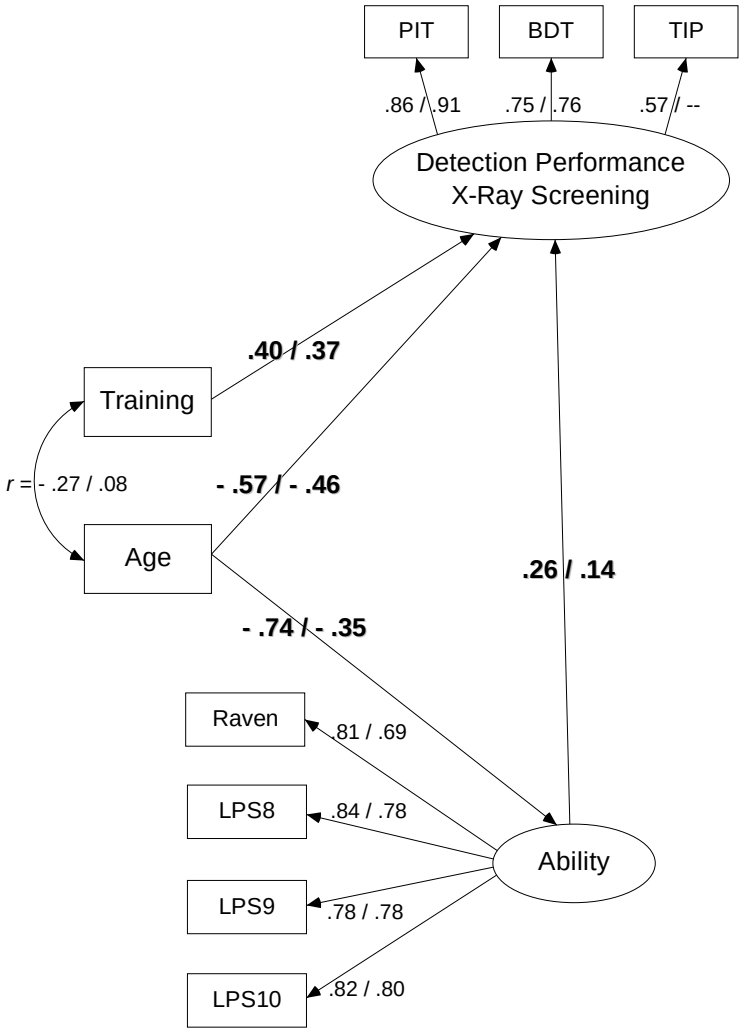

Figure 4.3. Direct and indirect effects (in bold type). Left for 2007 data (N = 97), right for the screener group (N = .130).

All direct, indirect and total effects are displayed in Table 3.7. Age shows the strongest direct effect in the model. Ability and detection performance is significantly lower for older employees -.74, $p < .01$ and -.57, $p < .01$, respectively. Further, results evidence that training influence the detection performance significantly .40, p

< .01. Ability has a positive effect on detection performance .26, p = .07. After controlling for age and training, ability is still positively associated with work performance. Furthermore, age has not only a direct, but also a significant indirect and negative effect on detection performance of -.19, p < .01. This results in a total effect of age on detection performance of -.76, p < .01.

Table 3.7

Direct, indirect and total effects for both samples

	standardized coefficients	direct	indirect	total
2007 sample (N=97)	Ability	.26	--	.26
	Training	.40	--	.40
	Age	-.57	-.19	-.76
Screener group (N=130)	Ability	.14	--	.14
	Training	.37	--	.37
	Age	-.46	-.05	-.51

This model could be successfully replicated with already employed aviation security screeners X^2(16, N= 130) = 17.78, p = .34, CFI = .996, RMSEA = .029, Pclose = .648, AIC = 57.78 (saturated model 72.00), BCC = 60.78 (saturated model 77.40), BIC = 115.13 (saturated model 175.23), CAIC = 135.13 (saturated model 211.23). For direct, indirect and total effects see Table 3.7.

In order to test whether the effect of training and ability varies across different age groups, a multiple group comparison across low and high age groups (median split) was conducted for the 2007 sample. However, results show no significant interaction of age with ability and training (p > .10). In this sample the effect of both ability and training remains unchanged for the different groups. However, it could be expected that samples with higher variance would show different results.

6.5 GENERAL DISCUSSION

The main goal of this study was to examine the relationship between different factors such as ability, personality traits, age etc. and detection performance in X-ray screening in order to define a reliable and valid pre-employment assessment. Factors

that are rather related to abilities and are important for the X-ray screening task were evaluated primarily. Results of Experiment 1 show that all cognition tests from the CTB which are mostly tests from elaborated German intelligence test batteries load on one latent factor ability despite their semantic distinctions. Furthermore, this factor correlates highly with detection performance in X-ray screening for both samples. A multiple group confirmatory factor analysis evidences the equality of factor loadings across both samples additionally. However, results indicated that one constraint (TIP) had to be released as this indicator was significantly different across the groups. Reliability of the 2006 sample which was just sufficient may account for this difference. Nevertheless we tried to integrate this indicator because of his importance to the latent variable detection performance in X-ray screening. Our results also suggest that the whole CTB which consists of 12 tests can be reduced to four tests without reducing explained variance. Further a full SEM with the X-Ray ORT as additional factor showed that both factors display a significant effect on detection performance. Interestingly, as well the X-Ray ORT which measures the ability to cope with image-based factors in X-ray screening seems to be an important determinant. This result was unexpected as participants in both groups were selected based on their test results in the X-Ray ORT and therefore variability of both groups is rather small. Further, an additional analysis showed that job applicants who performed high in the X-Ray ORT had as well significantly better test results in the CTB. To sum up Experiment 1 showed that both the ability measured with the X-Ray ORT and the ability measured with the CTB play an important role for the X-ray screening task later on the job. The positive relationship between the X-Ray ORT and detection performance later on the job could also be shown in a previous study by Hardmeier et al. (2006a). Thus, the X-Ray ORT as well as the CTB can be used within a pre-employment assessment. However, to increase efficiency a reduction of the CTB from 12 to 4 tests only should be taken into consideration.

In a next step all factors that can be relevant for the X-ray screening job and whereof reliable data are available were subjected to SEM. The model revealed that factor loadings for age and training are substantial. Thus, training hours increase the detection performance in both the competency assessment and on the job. This result is consistent with previous studies which showed that detection performance can be significantly improved with an individual adaptive training system (Hardmeier et al., 2006b; Koller et al., 2008). Further, a negative effect of age on detection performance could also be found by Schwaninger, Hardmeier, Riegelnig, and Martin

(in preparation). They could show that despite of their working experience older screeners performed significantly worse than their younger colleagues. Taking into consideration that the age range in this study was from 20 to 55 years only, this result is rather surprising. Further, ability has a weak and positive effect on detection performance. It could be expected that this effect would be stronger with larger sample size. Furthermore, it has to be considered that the sample used for this study shows relatively small variance as screeners were already selected based on their ability. Whether ability is even more important is a question that should be answered with a representative sample. Moreover, the four personality factors extraversion, conscientiousness, agreeableness and neuroticism show a rather weak, but positive effect on detection performance. It could be expected that these personality traits become even more important considering other tasks of aviation security screeners, such as the communication with passengers. Further analysis should investigate these relationships to decide if these personality factors should be measured within a pre-employment assessment. For working experience a negative effect on detection performance was found. However, the positive relationship between detection performance and working engagement alone implies a suppressor effect. In this study some indicator had to be neglected due to the small sample size. Further studies with larger sample size should investigate these effects more in detail. Last, an analysis about direct and indirect effects with the three most important factors age, training and ability showed that age influences detection performance directly, but also indirectly by a negative relationship with ability. Further, training and ability show a direct effect on the X-ray screening task only. This model could be replicated with another sample including aviation security screeners who were employed for at least four years.

Despite the limitation of this study due to the small sample size and many free parameters to be estimated factor loadings for ability, age and training were invariant over time. This evidences the importance of all three factors for the X-ray screening task. Whereas ability can be measured within a pre-employment assessment, training have to be provided when employed. In order to interpret the age effect reliably, further studies should be done. Although older screeners seem to perform significantly worse, a study by Schwaninger et al. (in preparation) could show that some older screener perform on remarkably higher level than their younger colleagues.

PART II

TRAINING IN AVIATION SECURITY

7. THE ROLE OF RECURRENT CBT FOR INCREASING AVIATION SECURITY SCREENERS' VISUAL KNOWLEDGE AND ABILITIES NEEDED IN X-RAY SCREENING

7.1 ABSTRACT

X-ray screeners have to know which items are prohibited and what they look like in X-ray images of passenger bags (knowledge-based factors). In addition, effective X-ray screening requires the abilities to cope with bag complexity, superposition by other objects, and rotation of objects (image-based factors). Knowledge-based factors are expected to be highly dependent on training whereas image-based factors are related to visual-cognitive abilities and aptitudes (Schwaninger et al., 2005).

To test to what extend these two factors are influenced by training, 334 screeners took two X-ray screening tests before and after two years of recurrent computer-based training (CBT). The Prohibited Items Test (PIT) measures rather knowledge-based factors, the X-Ray Object Recognition Test (X-Ray ORT) image-based factors. The results showed indeed a much better detection performance in the PIT after two years of training. Thus, CBT can increase the knowledge of prohibited items and what they look like in X-ray images of passenger bags substantially. The increase in detection performance in the X-Ray ORT was much smaller. This indicates that image-based factors are indeed related to visual-cognitive abilities and aptitudes that can be increased by CBT less effectively. The implications for selection and training of X-ray screeners are discussed[15].

7.2 INTRODUCTION

Airport security has become very important in recent years. Since airports are confronted with new threat dimensions and a constantly increasing passenger flow, reliable and efficient detection of different threat items in X-ray images is an essential airport security task. During high passenger flow, screeners have only a few seconds to decide whether a bag is OK or whether it has to be hand-searched. Schwaninger et al. (2005) could show that detecting threat items in passenger bags

[15] A slightly different version of this chapter was published at the 4th International Aviation Security Technology Symposium in Washington, D.C. 2006. I gratefully acknowledge the help of Adrian Schwaninger and Franziska Hofer in preparing the manuscript.

includes both, knowledge-based and image-based factors. That is, a screener has to know which items are prohibited and what they look like in X-ray images of passenger bags (knowledge-based factors). There are many threat items that look quite different in reality than in an X-ray image, which makes them difficult to recognize without training (Schwaninger, 2005b). Another difficulty results from the fact that some threat items look quite similar to harmless objects. Again other threat items like improvised explosive devices (IEDs) are normally not seen at checkpoints and without specific training they are therefore rather difficult to recognize. These knowledge-based factors are expected to be highly dependent on training since results from object recognition studies show that one can only recognize shapes if they are similar to the ones encountered before and stored in visual memory (for an overview see Schwaninger, 2005a). However, the interpretation of X-ray images is much more complex and is also dependent on image-based factors such as bag complexity, superposition and rotation of the target object itself. If a bag is close-packed it is more difficult to detect a threat item within a short time because other objects can distract attention. Furthermore, the threat item can be superimposed by other objects in the bag, which can hamper detection performance as well. Moreover, if a threat item is shown in a rotated view it becomes harder to recognize it. A previous study from Schwaninger et al. (2005) showed large individual differences for novices and screeners regarding how well they can cope with such image-based factors. Furthermore, it could be assumed that these image-based factors are less dependent on training and more related to visual abilities and aptitudes which are relatively stable over time.

To measure knowledge-based and image-based factors in X-ray screening relatively independent of each other, two X-ray screening tests, the Prohibited Items Test (PIT) and the X-Ray Object Recognition Test (X-Ray ORT) were developed. The PIT measures rather knowledge-based factors in X-ray screening and therefore includes all kinds of prohibited items according to international prohibited items lists. On the other hand the X-Ray ORT includes only guns and knives in X-ray images, but shown in different viewpoints with low and high superposition and in bags with different complexity levels and therefore measures mainly the image-based factors viewpoint, superposition and bag complexity.

To test to what extend knowledge-based and image-based factors are influenced by training, 334 aviation security screeners took both X-ray screening tests before and after two years of recurrent computer-based training with X-Ray Tutor, which is an

individually adaptive training system for X-ray screeners (see Schwaninger, 2004; Schwaninger & Hofer, 2004 for details).

7.3 METHOD

7.3.1 Participants

A total of 334 aviation security screeners (101 male and 233 female) between 23 and 62 years (*M* = 46.71, *SD* = 8.37) participated in this study. When taking the PIT and the X-Ray ORT the first time, all of them had a working experience between one and 25 years (*M* = 7.54, *SD* = 5.13). Between the first and the second measurement all screeners received two years of recurrent Computer Based Training (CBT) with X-Ray Tutor. Most screeners trained at least twice a week for 20 minutes.

7.3.2 Materials and Procedure

Prohibited Items Test (PIT)

The Prohibited Items Test (PIT) measures rather knowledge-based factors in X-ray image interpretation tasks and includes X-ray images of all kinds of prohibited items. All of them can be classified into the seven categories guns, sharp objects, hunt and blunt instruments, chemicals, highly inflammable substances, explosives and others according to ECAC, ICAO and EU prohibited items lists. The test contains a total of 160 X-ray images, half of them include one or more prohibited items, whereas the other 80 images are bags containing only harmless objects. 68 of the threat images contain exactly one prohibited item; the remaining 12 bags include two or three prohibited items[16]. As this test was developed to measure rather knowledge-based factors, all image-based factors are kept relatively constant. That is, all prohibited items were shown in a bag with medium bag complexity, medium superposition and in an easy view.

The PIT is a computer based test and includes a self-explanatory instruction with six exercise trials to familiarize the participants with the test taking procedure. All X-ray images are shown for a maximum of 10 seconds on the screen. Then, participants have to decide whether the bag is OK (includes no prohibited item) or NOT OK (includes one or more prohibited items) by clicking on the respective button.

[16] This was done to increase face validity of the test. Please note that only bags including one prohibited item were used for that data analysis.

Furthermore, they have to indicate how sure they are in their decision and to which of the seven categories the prohibited item belongs to[17]. The test is divided into four blocks of trials. After each block screeners had the possibility to take a short break. Trials are counterbalanced across all four blocks and the order within a block is random. The test takes about 45 minutes (short breaks included).

A previous study with 453 X-ray screeners could show that the PIT is a reliable and valid instrument to measure knowledge-based factors in X-ray screening (Hardmeier et al., 2006a). Reliability was measured using Cronbach Alpha and Guttman split half reliabilities; the former ranging between .87 and .93 and the latter ranging between .87 and .94. Furthermore, convergent, discriminant and criterion-related validity support high validity of the PIT (for more details see Hardmeier et al., 2006a).

X-Ray Object Recognition Test (X-Ray ORT)

The X-Ray Object Recognition Test (X-Ray ORT) measures rather image-based factors in X-ray screening and therefore includes only typical gun and knife shapes as threat objects. These are shown in bags with different complexity levels, more or less superimposed by other objects in the bag. Furthermore, each gun and each knife is shown in an easy and rotated view. Thus, the *X-Ray ORT* consists of a total of 256 trials: 2 threat categories (guns and knives) * 8 (exemplars) * 2 (bag complexities) * 2 (superpositions) * 2 (views) * 2 (harmless images vs. threat images). All images are shown in black and white so that this test can also be used for pre-employment assessment purposes where the meaning of color information as indicator for different materials is not known to novices.

The procedure is similar to the PIT. After a self-explanatory introduction, participants receive eight exercise trials with feedback. The test is also subdivided into four blocks. Trials are counterbalanced across all blocks and random within each block. Contrary to the PIT, images in the ORT are displayed for four seconds only and then participants answer whether the bag was OK or NOT OK. Again, at the end of each answer participants have to indicate how sure they are in their decision.

Detailed reliability and validity measures of this test can be found in Hardmeier et al. (2005). Overall, reliabilities with >.89 for Cronbach Alpha and >.78 for Guttman split half in the screener group are rather high. As well convergent, discriminant and criterion-related validity are given; for details see Hardmeier et al. (2005).

[17] Please note that data analysis is based on the OK/NOT OK answers and not on the answer to which category a prohibited item belongs to.

X-Ray Tutor – Individually adaptive computer based training system

X-Ray Tutor (XRT) is an individually adaptive training system to improve detection performance in X-ray screening. This CBT system creates individual training sessions adapted to each screener based on his learning history and thereby provides very effective and efficient training (for details see Schwaninger, 2004). XRT CBS 2.0 Professional Edition includes a large image library with 25'140 fictional threat item (FTI) images depicting more than 500 different threat objects in many different viewpoints. XRT combines each FTI with an X-ray image of a passenger bag during the training session in real-time. The training software starts with easy views at the beginning of the training. Depending on the screeners' learning history, image difficulty is increased by choosing more difficult viewpoints, increasing bag complexity and superposition adapted to each screener and FTI.

During a training session the X-ray image is displayed for 15 seconds on the screen. Then, screeners have to press an OK or NOT OK button to indicate whether the bag is harmless or whether it has to be hand-searched. Immediate feedback is provided, i.e. whether a screener has correctly identified (hit) or missed the threat item (miss), whether he/she correctly rejected a harmless bag (correct rejection) or wrongly judged a harmless bag as being dangerous (false alarm). Furthermore, an information window showing the X-ray image of the threat item and a real photograph of it provide immediate detailed information about the threat item and its components in order to enhance perceptual learning. The effectiveness of X-Ray Tutor has been proven in several scientific studies, showing substantial increases of detection, less false alarms and faster response times (Schwaninger & Hofer, 2004; Ghylin, Drury, & Schwaninger, 2006).

7.4 RESULTS

Test results are calculated using the detection performance measure d', which takes the hit rate and the false alarm rate into account. The hit rate shows how often a bag containing a threat item was judged as being not ok, whereas the false alarm rate shows how often a harmless bag was wrongly judged as not ok (Green & Sweets, 1966).

Figure 5.1a shows the difference in detection performance after two years of recurrent CBT for both tests, the PIT and X-Ray ORT. An analysis of variance (ANOVA) with the within-participant factors test type (PIT, ORT) and measurement

(first, second) was calculated using d' scores. There was a significant main effect of test type (PIT vs. ORT) $\eta^2 = .88$, $F(1, 333) = 2483.84$, $p < .001$, a significant main effect of measurement (first vs. second) $\eta^2 = .77$, $F(1, 333) = 1129.58$, $p < .001$ and a significant interaction of test type and measurement $\eta^2 = .34$, $F(1, 333) = 170.44$, $p < .001$. As can be seen, CBT had a much higher influence on detection performance in the PIT than in the X-Ray ORT.

Schwaninger et al. (2005) predicted a rather high training effect in the PIT and a small influence of training on image-based factors which depend mainly on visual abilities and aptitudes. The significant interaction between test type and measurement is consistent with this assumption and shows that image-based factors, i.e. the ability to cope with bag complexity, superposition and rotation of the threat item, can not be increased very much by training when compared to knowledge-based factors which depend highly on training.

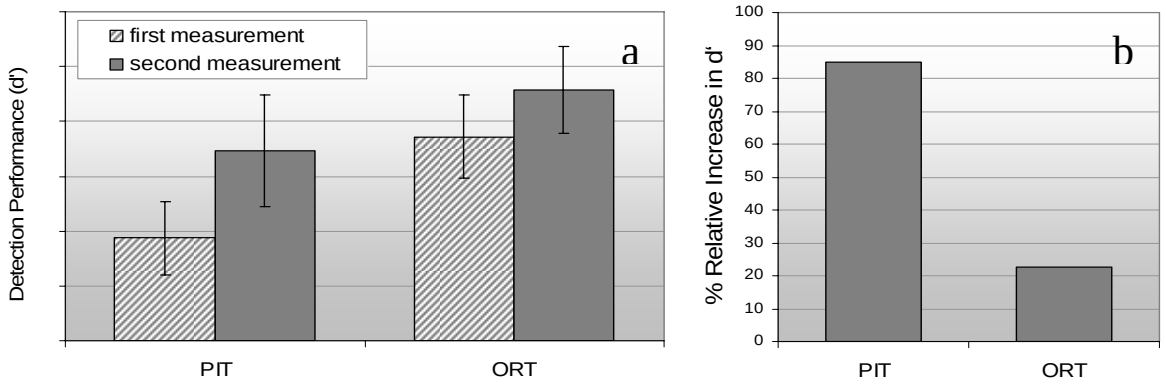

Figure 5.1. Detection Performance with standard deviations in the PIT and ORT for the first and second measurement left and percent difference in these two tests right.

Figure 5.1b shows the difference in percentage for both tests, taking the first measurement as baseline using the following formula (m1 means first measurement and m2 means second measurement):

$$\%\ \text{Relative Increase in d'} = \left(\frac{m2 - m1}{m1}\right) * 100$$

As predicted by Schwaninger et al. (2005), there were large effects of training on knowledge-based factors (85.0%) and rather low influence of training on image-based factors (22.7%).

7.5 DISCUSSION

The detection of threat items in passenger bags depends on knowledge-based and image-based factors. Screeners have to know which objects are prohibited and what they look like in X-ray images in order to recognize them (knowledge-based factors). In addition, they have to be able to cope with effects of bag complexity, superposition, and viewpoint (image-based factors).

In this study we investigated the role of training on knowledge-based and image-based factors in aviation security screening using the two X-ray screening tests, PIT and X-Ray ORT as well as XRT, a computer based individually adaptive training system. Overall, results show that the increase in detection performance after two years of recurrent computer-based training was much smaller in the X-Ray ORT compared to the training effect in the PIT. These results support the assumption that the PIT measures rather knowledge-based factors which can be increased remarkably through training compared to image-based factors measured by the X-Ray ORT, which are more difficult to increase by training. These results are as well consistent with results from a previous study by Schwaninger et al. (2005) which could show that the difference in X-ray detection performance between novices and aviation security screeners is much higher in the PIT than in the X-Ray ORT.

As can be seen in Figure 5.1a, detection performance in the PIT increased remarkably after two years of recurrent computer based training. Thus, an individually adaptive computer based training system like XRT helps to learn and store all kinds of prohibited items in different views in the visual memory. Additionally, it provides an excellent tool to react immediately to new threats as the training library of XRT can be updated constantly in an easy and fast way. However, besides the knowledge of a screener the abilities to cope with the image-based factors viewpoint, superposition and bag complexity are also very important. Hardmeier et al. (2005) revealed large inter-individual differences regarding the visual abilities needed to cope with these image-based factors, which affected on the job performance. As this study shows, these abilities can only be trained to a limited extent and therefore should be measured in a pre-employment assessment to identify the candidates who are well-suited regarding these visual abilities.

To summarize, both knowledge-based and image-based factors are very important in X-ray screening and can be measured relatively independent of each other using the PIT and X-Ray ORT. While knowledge-based factors can be enhanced remarkably by

adaptive computer-based training, image-based factors should already be measured and used for selecting candidates in pre-employment assessment.

8. INVESTIGATING TRAINING, TRANSFER AND VIEWPOINT EFFECTS RESULTING FROM RECURRENT CBT OF X-RAY IMAGE INTERPRETATION

8.1 ABSTRACT

X-ray screening of passenger bags is an essential task at airport security checkpoints. In this study we investigated how well airport security screeners can detect guns, knives, improvised explosive devices (IEDs) and other threat objects in X-ray images of passenger bags before and after three and six month of recurrent (about 20 min per week) computer-based training (CBT). Two experiments conducted at different airports gave very similar results. Training with X-Ray Tutor (XRT), an individually adaptive CBT, resulted in large performance increases, especially for detecting IEDs. While performance for detecting IEDs was initially substantially lower than for guns, IEDs could be detected as well as guns after several months of training. A large transfer effect was observed as well: Training with XRT helped screeners recognize new threat objects that were similar in shape as the trained objects. Threat recognition was dependent on the rotation of the objects. If depicted from an unusual viewpoint, prohibited items were more difficult to recognize. The results were compared to two conventional CBT systems. For one system no training and transfer effects were observed whereas small training and transfer effects were found for the other conventional CBT system[18].

8.2 INTRODUCTION

The importance of aviation security has increased dramatically in the last years. As a consequence of the new threat situation, large investments were made into modern security technology. State of the art X-ray screening equipment offers good image quality, high resolution and many image enhancement functions. However, the decision whether an X-ray image of a passenger bag contains a prohibited item or not, is still being taken by a human operator, i.e. an airport security screener. Object shapes that are not similar to ones stored in visual memory are difficult to recognize

[18] I gratefully acknowledge the help of Saskia Koller, Adrian Schwaninger and Stefan Michel in conducting the study and preparing the manuscript. A similar version was published in the Journal of Transportation Security, 2008.

(e.g., Graf et al., 2002; Schwaninger, 2004, 2005a). Schwaninger et al. (2005) have shown that X-ray screener performance depends on knowledge-based and image-based factors. A prerequisite for good X-ray detection performance is knowledge about which objects are prohibited and what they look like in X-ray images. Such knowledge is acquired by computer-based, class-room and on the job training (knowledge-based factors). Image-based factors refer to image difficulty resulting from viewpoint variation of threat objects, superposition of threat objects by other objects in a bag, and bag complexity depending on the number and type of other objects in the bag. The ability to cope with image-based factors is related to individual visual-cognitive abilities rather than a mere result of training (Hardmeier et al., 2006b). Because many threat objects are not known from everyday experience and because objects look quite different in X-ray images than in reality computer based and on the job training, as well as job experience are expected to be important determinants of X-ray detection performance. This is illustrated in Figure 1.2. Improvised explosive devices (IEDs) are normally neither seen at checkpoints nor in reality and therefore very difficult to recognize for untrained persons. Schwaninger and Hofer (2004) and Schwaninger, Hofer, and Wetter (2007) could show that detection of IEDs in hold baggage screening (HBS) can be significantly improved if people are trained with an individually adaptive training system. Thus, training helps to store unknown object shapes in visual memory. Furthermore, some threat items, e.g. a gas spray looks quite different in the X-ray image. Schwaninger et al. (2005) compared detection performance of novices with the one of aviation security screeners. A rather poor recognition of unfamiliar object shapes (e.g., self-defense gas spray, electric shock device etc.) in X-ray images was found for novices. For experienced aviation security personnel, a much higher recognition performance was observed. Consistent with this result McCarley, Kramer, Wickens, Vidoni, and Boot (2004) reported a better performance after training for the detection of knives in X-ray images for novices. However, one could assume that the expertise for other threat categories than IEDs which are relatively often seen at checkpoints (such as knives) can be gained with job experience and on the job training alone and that individually adaptive training is far less important for these threat categories.

When one takes into account the myriad of views that can be produced by a single object, the question arises how the human brain stores and recognizes objects even if they are presented in unusual views. In the object recognition literature, two types of theories can be distinguished: structural description theories and view-based

theories. The former assume that objects are stored in visual memory by their component parts and their spatial relationship. An objects-centered description of this nature was described by Marr and Nishihara (1978), who proposed that objects are hierarchically decomposed into their parts and spatial relations relative to object-centered coordinates in order to access an object-centered 3D model in visual memory. In Biederman´s (1987) recognition by components (RBC) theory, non-accidental properties like vertices, parallel vs. non-parallel lines, straight vs. curved lines etc. (see Lowe, 1985, 1987) are extracted from a line drawing representation of objects to define basic geometrical primitives (geometrical ions, "geons") that are relatively orientation-invariant. A geon structural description (GSD) in memory is activated by extracting geons from the visual input and match geon properties and their spatial relationship with the GSD (Hummel & Biederman, 1992).

For view-based theories, different approaches have been proposed. Examples are recognition by alignment to a 3D representation (Lowe, 1987), recognition by linear combination of 2D views (Ullman & Basri, 1991), recognition by view interpolation (e.g., using RBF networks) proposed by Poggio and Edelman (1990) and storing of multiple views for each object plus performing transformations (Tarr & Pinker, 1989). What view-based theories have in common is the assumption that objects are not stored in memory as rotation invariant structural descriptions but instead in a format which is viewer-centred. A more detailed discussion of structural description theories vs. view-based theories and more recent hybrid theories is beyond the scope of this paper (for reviews see for example Graf et al., 2002; Hayward, 2003; Kosslyn, 1994; Peissig & Tarr, 2007; Schwaninger, 2005a; Tarr & Bülthoff, 1998). However, it should be pointed out that empirical results seem to be correlated with the required level of recognition (Bülthoff, Edelman, & Tarr, 1995; Tarr, 1995): if the object has to be recognized at 'entry level', behavioral measures are less affected by changes in perspective. However, in the case of subordinate recognition in which fine discriminations are typically required, both response times and accuracy are far more sensitive to the specific viewpoint used. Furthermore, differences in the task a subject has to perform (Lawson, 1999) and the specific paradigm that is used (Verfaillie, 1992) can influence which level of representation is tapped (see also Logothetis & Sheinberg, 1996).

The first aim of this study is to investigate how well airport security screeners can detect guns, knives, IEDs and other prohibited items in X-ray images of passenger bags. The second aim is to examine whether screener detection performance can be

increased by conducting recurrent CBT and whether this increase can be shown for all threat categories. To this end, screeners conducted weekly recurrent CBT (about 20 min per week). Detection performance was tested with the X-Ray Competency Assessment Test (X-Ray CAT) by Koller and Schwaninger (2006). This test measures how well people detect threat items in X-ray images of passenger bags. It was conducted at the beginning and then after three and six months of training. In addition to training effects, The X-Ray CAT allows measuring transfer effects, i.e. to what extent visual knowledge that was gained through CBT can be transferred to other threat items (see below). In the X-Ray CAT all prohibited items are depicted from a canonical (easy recognizable) perspective (Palmer et al., 1981) and unusual perspective which allows investigating viewpoint effects. The study was conducted at two mid-size European airports. In Airport 1 (Experiment 1) one group of screeners used adaptive CBT (XRT) whereas the other group of screeners (control group) used a conventional (not adaptive) CBT. In Airport 2 (Experiment 2) the same experimental design was used except for the fact that the control group used another conventional CBT system. This allows finding out if a training effect is as well dependent on the type of the CBT system used.

8.3 EXPERIMENT 1

8.3.1 Method

8.3.1.1 Participants

209 airport security screeners of a mid-size European airport participated in Experiment 1 and conducted the X-Ray CAT 1.0.0 three times with an interval of three months between the measurements. The adaptive CBT group (XRT group) consisted of 97 screeners who conducted weekly recurrent CBT using X-Ray Tutor (XRT) CBS 2.0 Standard Edition between all three test measurements. The control group consisted of 112 screeners who used a conventional (not adaptive) CBT. According to the security organization and their Appropriate Authority, airport security screeners of both groups conducted about 20 min CBT per week. Analysis of XRT training use showed that on average, each screener trained 20.26 minutes (*SD* = 3.65 min) per week.

8.3.1.2 Materials

The X-Ray Competency Assessment Test (X-Ray CAT)

The X-Ray CAT consists of 256 trials based on 128 different X-ray images of passenger bags. Each of the bag images is used once containing a prohibited item (threat image) and once without any threat object (non threat image). Figure 6.1 displays examples of the stimuli.

Figure 6.1. Example images from the X-Ray CAT. Left: harmless bag (non threat image), right: same bag with a prohibited item at the top right corner (threat image). The prohibited item (gun) is shown also separately at the bottom right.

Prohibited objects can be assigned to four categories as defined in Doc 30 of the European Civil Aviation Conference (ECAC): guns, IEDs, knives and other prohibited items (e.g., gas, chemicals, grenades etc.). The threat objects have been selected and prepared in collaboration with experts of Zurich State Police, Airport Division to be representative and realistic. For each threat category 16 exemplars are used (eight pairs). Each pair consists of two prohibited items that are similar in shape (see Figure 6.2). These were distributed randomly into two sets, set A and set B. Prohibited items of set A (not threat bag images) are contained in the XRT CBS 2.0 SE training whereas the items of set B are not. This allows testing for transfer effects.

Every item is depicted from two different viewpoints. The easy viewpoint refers to the canonical (i.e. easy recognizable) perspective (Palmer et al., 1981). The difficult viewpoint shows the threat item with an 85 degree horizontal rotation or an 85 degree vertical rotation relative to the canonical view (see Figure 6.2 for examples). In each threat category, half of the prohibited items of the difficult viewpoint are rotated vertically, the other half horizontally.

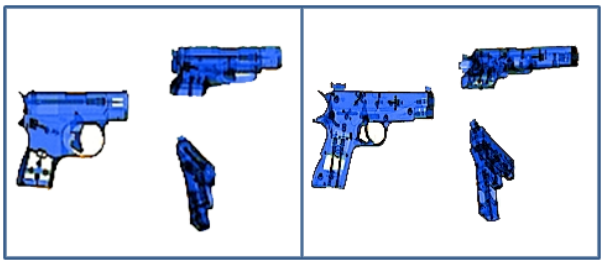

Figure 6.2. Example of two x-ray images of similar looking threat objects used in the test. Left: a gun of set A. Right: Corresponding gun of set B.

Set A and B are equalized concerning the rotations of the prohibited objects.
Every threat item is combined with a bag in a manner that the degree of superposition by other objects is similar for both viewpoints. This was achieved using a function that calculates the difference between the pixel intensity values of the bag image with the threat object minus the bag image without the threat object using the following formula:

$$SP = \frac{\sqrt{\sum \left[I_{SN}(x,y) - I_{N}(x,y) \right]^2}}{ObjectSize}$$

SP = Superposition; I_{SN} = Grayscale intensity of the SN (Signal plus Noise) image (contains a prohibited item); I_N = Grayscale intensity of the N (Noise) image (contains no prohibited item); Object Size: Number of pixels of the prohibited item where R, G and B are < 253

Using this equation (division by object size), the superposition value is independent of the size of the prohibited item. This value can be kept relatively constant for the two views of a threat object, independent of the degree of clutter in a bag, when combining the bag image and the prohibited item. The bag images were visually inspected by aviation security experts to ensure they do not contain any other prohibited items. Harmless bags were assigned to the different categories and viewpoints of the threat objects in a way that their difficulty was balanced across all categories[19]. The false alarm rate (the rate at which screeners wrongly judged a harmless bag as containing a threat item) for each bag image served as measure of difficulty based on a pilot study with 192 screeners of another airport.
The X-Ray CAT takes about 20 to 30 minutes to complete. Each image is shown for a maximum of 10 seconds on the screen. Screeners have to judge whether the bag is OK (contains no prohibited item) or NOT OK (contains a prohibited item). Additionally, screeners have to indicate the perceived difficulty of each image on a 100 point scale (difficulty rating)[20]. The X-Ray CAT is built into the XRT training system (see below). The interface of the X-Ray CAT is the same as in XRT except there is no feedback and screeners do not have to click on the image to identify the threat object.

[19] The eight categories of test images (four threat categories in two viewpoints each) are similar in terms of the difficulty of the harmless bags. This means, a difference of detection performance between categories or viewpoints can not be due to differences in the difficulty of the bag images.
[20] The difficulty ratings were not analyzed in study.

The X-Ray Tutor (XRT) Training System

X-Ray Tutor (XRT) is an individually adaptive training system for aviation security screeners. It contains a large image library with 400 different threat objects depicted

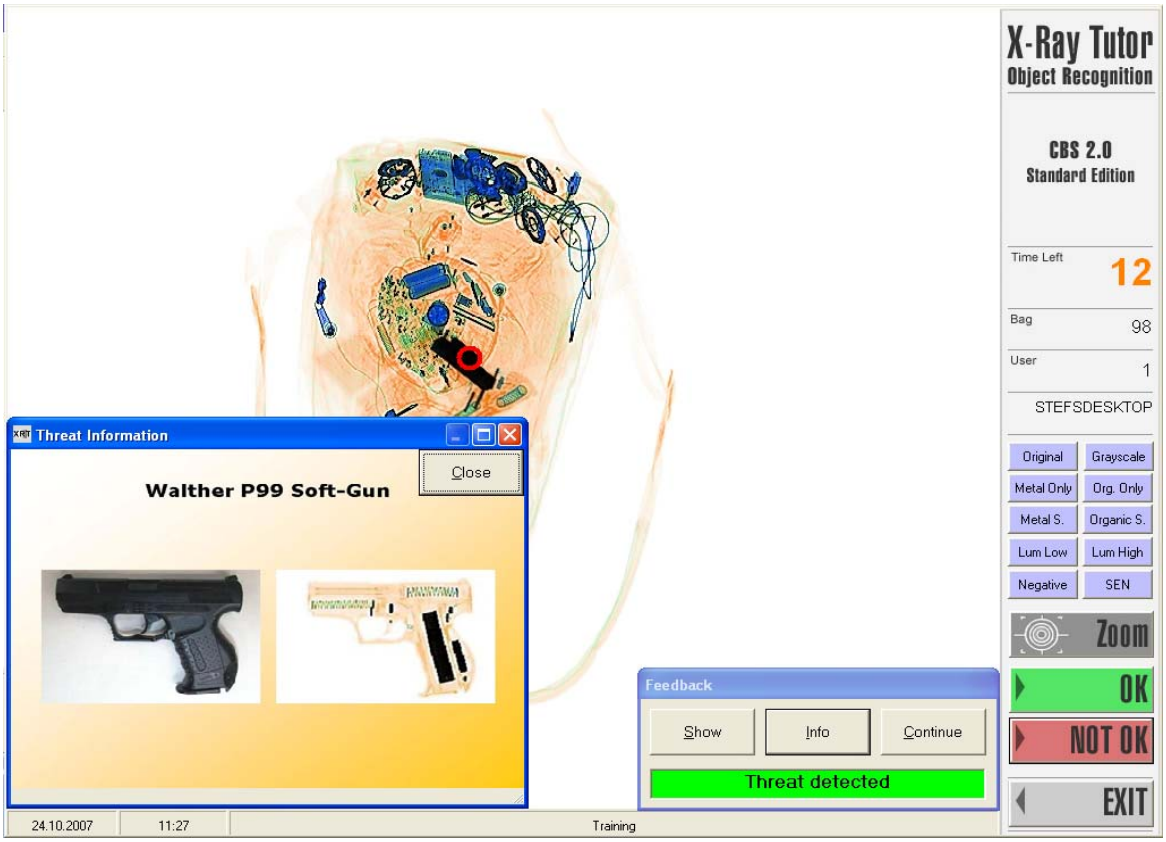

Figure 6.3. Screenshot of the XRT CBS 2.0 training system during training. At the bottom right a feedback is provided after each response. If a bag contains a prohibited item, an information window can be displayed (see bottom left of the screen).

in up to 72 views, more than 6000 bag images and millions of possible threat object to bag combinations (see Schwaninger 2004b for details). The individually adaptive training algorithm of XRT starts with showing threat objects depicted from easy viewpoints with little superposition by other objects in bags of little complexity. Based on each individual screeners´ learning progress, threat objects are shown in more difficult views, more complex bags and with more superposition. These parameters are adapted automatically by a scientifically validated algorithm for each screener and threat object while taking into account automatic image processing algorithms as explained in Schwaninger, Michel and Bolfing (2007). XRT first presents screeners prohibited objects in easy (canonical) views. The individually adaptive training algorithm determines for each screener which views are difficult to recognize and adapts the training so that the trainee becomes able to detect threat

items reliably even if prohibited objects are substantially rotated away from the easiest view. During the next difficulty levels, first superposition and then bag complexity is increased so that the trainee becomes able to detect threat items reliably even if they are superimposed by other objects or if the complexity of a bag is very high (for more information on XRT see Schwaninger, 2003b, 2004b, 2005c).
During a training session each image is displayed for 15 seconds on the screen. Within this time screeners can use some image enhancement functions which are also available when working with the X-ray machine (e.g. grayscale, negative image, edge enhancement, etc.). If the image contains a prohibited item, screeners have to click on it and then click on the NOT OK button. If the bag is harmless they have to click on the OK button. After providing a confidence rating using a slider control, feedback is shown to inform the trainee whether the image has been judged correctly or not (see Figure 6.3). If the bag contains a threat item, it is highlighted by flickering and the trainee has the possibility to display information about the threat item (see bottom left of Figure 6.3). By clicking on the continue button the next image is shown. As a default setting, one training sessions takes 20 minutes. During this time screeners see between 150 and 300 images.

8.3.1.3 Procedure

As explained above, two groups of screeners participated in Experiment 1. The XRT training group conducted weekly recurrent CBT using XRT CBS 2.0 Standard Edition. The control group used a conventional (not adaptive) CBT which is also used at many airports worldwide. Compared to the XRT it contains a smaller threat image library and threat objects are not displayed in many different views. Furthermore, in the conventional CBT threat objects are not matched with different bags on the fly and there is no individually adaptive training algorithm.
The XRT training group and the control group took the X-Ray CAT before, after three, and after six months of weekly CBT. This allows testing the effectiveness of both CBT systems for increasing X-ray image interpretation competency of airport security screeners. As explained above, half of the prohibited items in the X-Ray CAT are also contained in the XRT training system (although presented in different bags). The other half of the prohibited items of the X-Ray CAT is not part of the XRT training library. This allows testing for transfer effects, i.e. testing whether training with the detection of certain prohibited items helps increasing the detection of other prohibited items. Finally, as specified above in the section on the X-Ray CAT, all

prohibited items are depicted in easy and difficult view which allows testing effects of viewpoint on screener detection performance.

8.3.2 Results and Discussion

Detection performance was calculated using the signal detection measure d' (Green & Swets, 1966), which takes into account the hit rate (correctly judged threat images as being NOT OK) and the false alarm rate (wrongly judged harmless bags as being NOT OK). D' is calculated using the following formula:

$d' = z(H) - z(FA)$

H is the hit rate, FA the false alarm rate and z refers to the z-transformation. Performance values are not reported due to security reasons. However, effect sizes are reported for all relevant analyses and interpreted based on Cohen (1988), see Table 4.1. For *t*-tests, *d* between 0.20 and 0.49 represents small effect size; *d* between 0.50 and 0.79 represents medium effect size; $d \geq 0.80$ represents large effect size. For analysis of variance (ANOVA) statistics, η^2 between 0.01 and 0.05 represents small effect size; η^2 between 0.06 and 0.13 represents medium effect size; $\eta^2 \geq 0.14$ represents large effect size.

Table 4.1

Classification of effect sizes based on Cohen (1988)

Effect size	d	η^2
small	0.20-0.49	0.01-0.05
medium	0.50-0.79	0.06-0.13
large	≥ 0.8	≥ 0.14

Figure 6.5 shows the detection performance of the first, second and third measurement for both screener groups. As can be seen, there was a large improvement as a result of training in the XRT training group while there was no improvement in the control group. These results were confirmed by an ANOVA for repeated measures using d' scores with the within-participant factor measurement (first, second and third) and the between-participant factor group (XRT training group and control group). There were large main effects of measurement, $\eta^2 = .28$, $F(2, 414) = 81.04$, $p < .001$, and group, $\eta^2 = .19$, $F(1, 207) = 47.62$, $p < .001$. There was also a large interaction of measurement and group, $\eta^2 = .25$, $F(2, 414) = 68.67$, $p < .001$, which is consistent with Figure 6.5 showing large performance

increases as a result of training for the XRT training group but not for the control group.

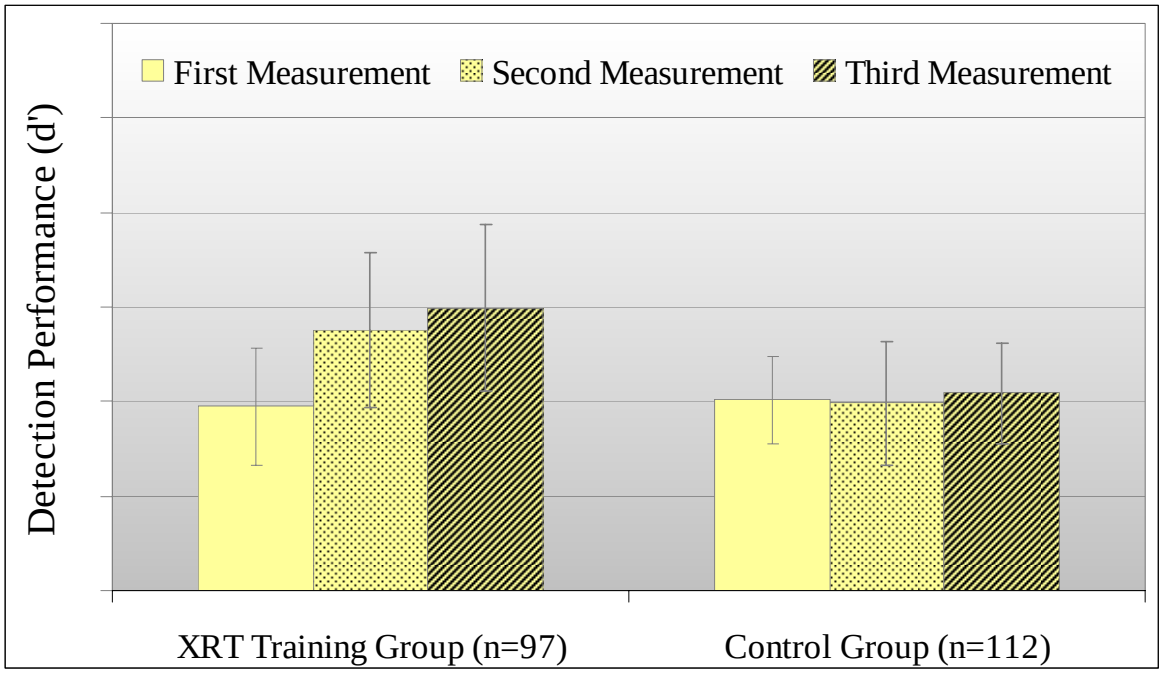

Figure 6.5. Detection performance with standard deviations for the XRT training group (left) vs. the control group (right) comparing first, second and third measurement.

Separate pairwise *t*-tests of detection performance d' revealed no significant difference at the baseline measurement between the two groups $t(177) = -0.91$, $p = .363$, $d = 0.13$, but already a significant difference in the second measurement, i.e. after three months of training, $t(207) = 7.52$, $p < .001$, $d = 1.04$. Additional paired-samples *t*-tests revealed significant differences for the XRT training group between all three test measurements but no significant differences for the control group (see Table 4.2).

Table 4.2

Results of the t-tests comparing the detection performance of first (t1), second (t2) and third (t3) measurement

	***t*(96)**	***p***	***d***
XRT Training Group (t1 – t2)	-9.80	< .001	1.12
XRT Training Group (t2 – t3)	-3.95	< .001	0.28
	***t*(111)**	***p***	***d***
Control Group (t1 – t2)	.54	= .59	0.05
Control Group (t2 – t3)	-1.89	= .06	0.17

Figure 6.6 shows the detection performance of both screener groups broken up by prohibited item category and the three test measurements. A repeated-measures ANOVA with the within-participant factors measurement (first, second and third) and threat category (guns, IEDs, knives and other), and the between-participant factor group (XRT training vs. control) revealed the significant main effects and significant interactions given in Table 4.3a. In addition to the effects that were already found in the previous ANOVA, also the factor threat (or prohibited item) category was significant. As can be seen in Figure 6.6, guns were detected best, followed by knives, other prohibited items and IEDs at the first test measurement. There was a highly significant interaction between threat category and measurement. As can be seen in Figure 6.6, detection of IEDs was initially much lower than gun detection. After six months of training, screeners in the XRT training group could detect IEDs even slightly better than guns. This result implies that IED detection is not difficult per se but rather a matter of knowledge which could be gained with specific training. Note that in this study all IEDs contained a detonator, wires, explosive, a triggering device and a power source. Thus, our conclusions are only applicable to the detection of such multi-component IEDs. Large performance increases were also found for other prohibited items in this group. Detection performance for guns and knives, threat items which are relatively often seen at checkpoints, was quite high in the first measurement condition already for both groups. This supports the assumption that job experience helps to store object shapes and increases detection performance to some amount. Nevertheless, for guns a remarkably improvement as a result of training can be found while for knives only a small training effect was shown. Note that after six months of training, detection performance of knives is lower than the one for any other threat category in the XRT training group, although at baseline

measurement it was higher than the detection performance for IEDs or other threat objects. The interaction between threat category, group and measurement is also worth mentioning. As can be seen in Figure 6.6 this results from the fact that there was no training effect for the control group. Their detection performance remains at about the same level for each threat category even after six months of training with the conventional CBT system.

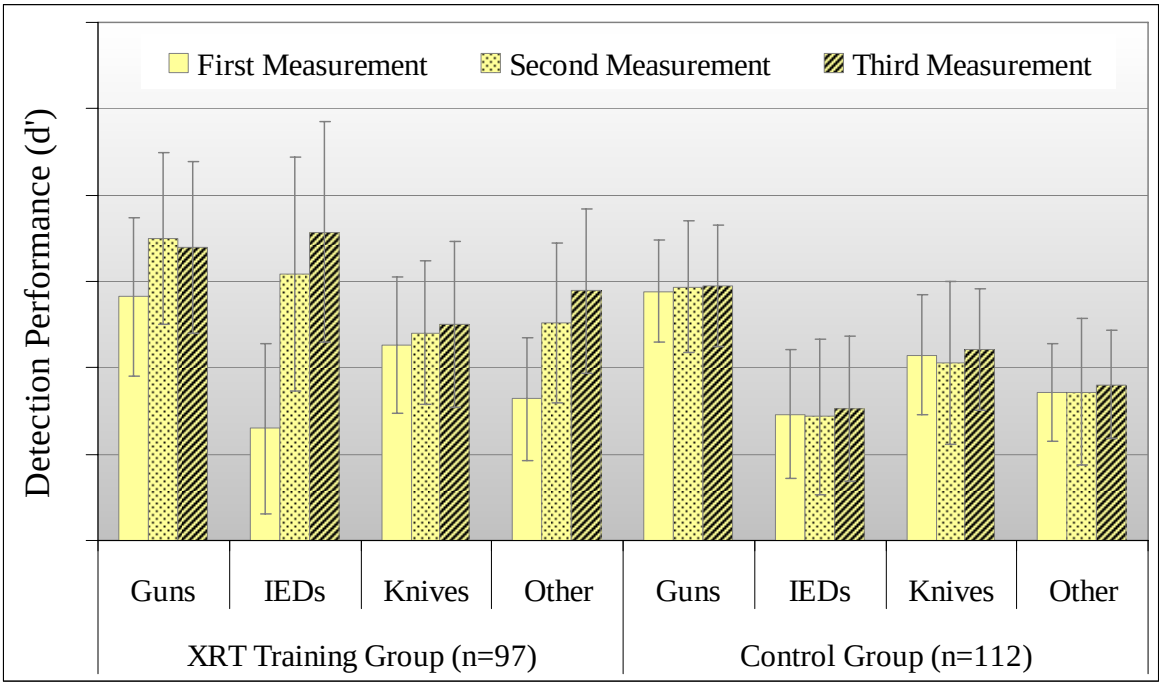

Figure 6.6. Detection performance with standard deviations for the XRT training group vs. the control group broken up by prohibited item category and test measurement.

Separate pairwise *t*-tests were conducted to compare detection performance at the first and the second measurement for both groups and each threat category separately (Table 4.4). The XRT training group showed a significant increase of the detection performance at the second measurement for the categories guns, IEDs and other threat objects. For knives, a significant difference could be found only in the third measurement. The comparison of the effect size *d* between the *t*-tests of the four threat categories confirms the earlier mentioned conclusion that the training effect was particularly big for IEDs and rather small for knives. Detection performance of the control group did not differ significantly between the measurements, confirming that the conventional CBT did not result in an increase of threat detection performance.

Table 4.3

Results of the ANOVAs in Experiment 1

	Factor	df	F	η^2	p
	Measurement (M)	2, 414	83.96	.29	< .001
	Threat Category (T)	3, 621	240.03	.54	< .001
	Group (G)	1, 207	56.20	.21	< .001
a)	M x G	2, 414	70.49	.25	< .001
	T x G	3, 621	45.05	.18	< .001
	M x T	6, 1242	43.20	.17	< .001
	M x T x G	6, 1242	40.65	.16	< .001
	Measurement (M)	2, 414	80.55	.28	< .001
	Set (S)	1, 207	4.18	.02	< .05
	Group (G)	1, 207	49.40	.19	< .001
b)	M x G	2, 414	67.99	.25	< .001
	M x S	2, 414	8.80	.04	< .001
	S x G	1, 207	51.32	.20	< .001
	M x S x G	2, 414	11.54	.05	< .001
	Measurement (M)	2, 414	87.69	.30	< .001
	Set (S)	1, 207	2.37	.01	= .13
	Threat Category (T)	3, 621	236.79	.53	< .001
	Group (G)	1, 207	63.57	.24	< .001
	M x G	2,414	71.16	.26	< .001
	M x T	6, 1242	44.35	.18	< .001
	M x S	2, 414	10.93	.05	< .001
c)	S x G	1, 207	52.25	.20	< .001
	S x T	3, 621	74.00	.26	< .001
	T x G	3, 621	47.39	.19	< .001
	M x T x G	6, 1242	41.04	.17	< .001
	M x S x G	2, 414	10.74	.05	< .001
	M x S x T	6, 1242	3.84	.02	< .01
	S x T x G	3, 621	4.78	.02	< .01
	M x S x T x G	6, 1242	2.99	.01	< .01

	Factor	df	F	η^2	p
	Measurement (M)	2,414	84.10	.29	< .001
	View (V)	1, 207	1768.63	.90	< .001
	Threat Category (T)	3, 621	258.62	.56	< .001
	Group (G)	1, 207	61.91	.23	< .001
	M x G	2, 414	65.80	.24	< .001
	M x T	6, 1242	41.33	.17	< .001
	M x V	2, 414	2.05	.01	= .13
d)	V x G	1, 207	3.27	.02	= .07
	V x T	3, 621	425.64	.67	< .001
	T x G	3, 621	40.86	.17	< .001
	M x T x G	6, 1242	40.25	.16	< .001
	M x V x G	2, 414	2.23	.01	< .05
	M x V x T	6, 1242	6.58	.03	< .001
	V x T x G	3, 621	3.08	.02	< .05
	M x V x T x G	6, 1242	2.68	.01	< .05

Table 4.4

Results of the t-tests comparing the detection performance of the four categories between the first (t1), second (t2) and third (t3) measurement

XRT training group	*t*(96)	df	*p*	*d*
Guns t1 – t2	- 5.96	96	< .001	0.70
IEDs t1 – t2	- 13.03	96	< .001	1.53
Knives t1 – t2	- 1.51	96	= .13	0.17
Other t1 – t2	- 8.47	96	< .001	1.07
Guns t1 – t3	- 4.69	96	< .001	0.60
IEDs t1 – t3	- 15.88	96	< .001	2.00
Knives t1 – t3	- 2.27	96	< .05	0.26
Other t1 – t3	- 12.56	96	< .001	1.51

Control group	***t*(111)**	**df**	***p***	***d***
Guns t1 – t2	- 0.40	111	= .69	0.05
IEDs t1 – t2	0.03	111	= .98	0.00
Knives t1 – t2	0.83	111	= .41	0.09
Other t1 – t2	-0.17	111	= .87	0.02
Guns t1 – t3	-0.92	111	= .36	0.10
IEDs t1 – t3	-1.05	111	= .30	0.08
Knives t1 – t3	-0.73	111	= .47	0.08
Other t1 – t3	-1.39	111	= .17	0.15

The results of the analyses considering the two prohibited item sets of the X-Ray CAT, set A and set B, are shown in Figures 6.7 and 6.8. As explained above, set A are X-Ray CAT images which contain prohibited items which are part of the XRT image library. Set B are X-Ray CAT images which contain prohibited items that are not part of the XRT image library. By comparing training effects for set A and set B transfer effects can be investigated, i.e. whether training with XRT does not only improve detection of prohibited items that are part of the XRT image library (set A) but also the detection of other prohibited items that are visually similar (set B). Figure 6.7 shows the detection performance for both screener groups broken up by test set for all three measurements. It shows a clear increase in detection performance for the XRT training group, especially at the second measurement, after the first three months of training. For the control group, as in the previous analysis, no training effect is evident. The results of the repeated measures ANOVA with the within-participant factors measurement (first, second and third) and set (A vs. B) and the between-participant factor group (XRT training group vs. control group) can be seen in Table 4.3b. There was a significant effect of set in this analysis, which would imply a different detection performance for set A vs. set B. However, the effect is very small, as the effect size of $\eta^2 = 0.2$ clearly shows, which makes the difference quasi negligible. This is also supported by the small effect size for the interaction between set and measurement, $\eta^2 = 0.4$. Pairwise *t*-tests comparing both sets within one group at the first measurement revealed a significant difference of

the two sets only for the control group $t(111) = -2.82$, $p < .01$, $d = 0.17$ but not for the XRT training group, $t(96) = -0.42$, $p = .68$, $d = 0.03$. However, note that an effect size of $d = 0.17$ is very small which supports the assumption that the two sets are in fact very similar in their difficulty level. Pairwise *t*-tests showed a significant increase in detection performance at the second measurement for both sets for the XRT training group, set A, $t(96) = -10.27$, $p < .001$, $d = 1.19$, set B, $t(96) = -7.68$, $p < .001$, $d = 0.92$. These results indicate a large transfer effect, i.e. visual knowledge regarding the visual appearance of the prohibited objects of the XRT image library helped screeners to detect similar looking, but untrained objects in the X-Ray CAT (set B). Consistent with previous analyses, there was no training effect for the control group, neither for set A, $t(111) = .76$, $p = .45$, $d = 0.08$, nor for set B, $t(111) = -0.28$, $p = .78$, $d = 0.03$.

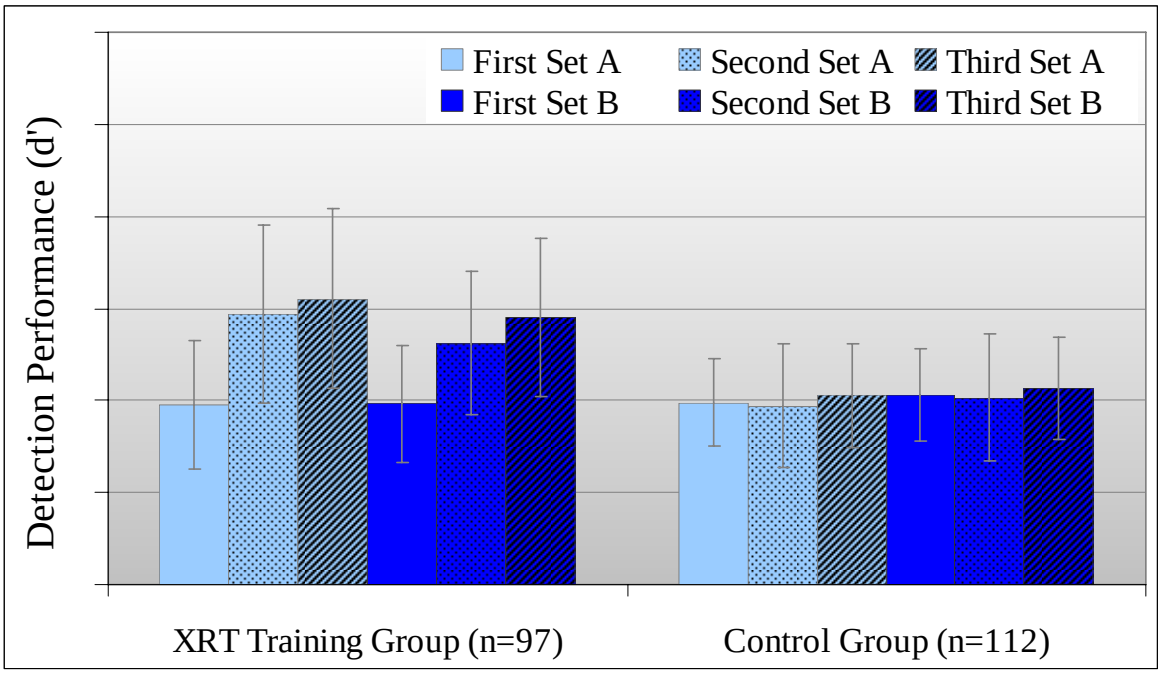

Figure 6.7. Detection performance with standard deviations for the XRT training group vs. the control group comparing first, second and third measurement for set A and set B separately.

Figure 6.8 includes also the threat category in the analysis. The increase in detection performance for the XRT training group can also be seen for the different threat categories. Pairwise *t*-tests between the first and second measurement confirmed a significant ($p < 001$, all $d > 0.62$) increase in detection performance for the XRT training group for all threat categories per set except for knives (set A: $p = .12$, $d = 0.19$, set B; $p = .32$, $d = 0.12$). In Figure 6.8, detection performance in Set A for guns shows a decrease between the second and third measurement. However, this

difference was not significant (p = .13, d = 0.17). For the control group, detection performance between the first and *third* measurement was compared in order to maximize the chances for finding a significant training effect. Even here, for all categories in each set, the detection between the first and third measurement did not differ significantly (all p > .12, d < 0.18).

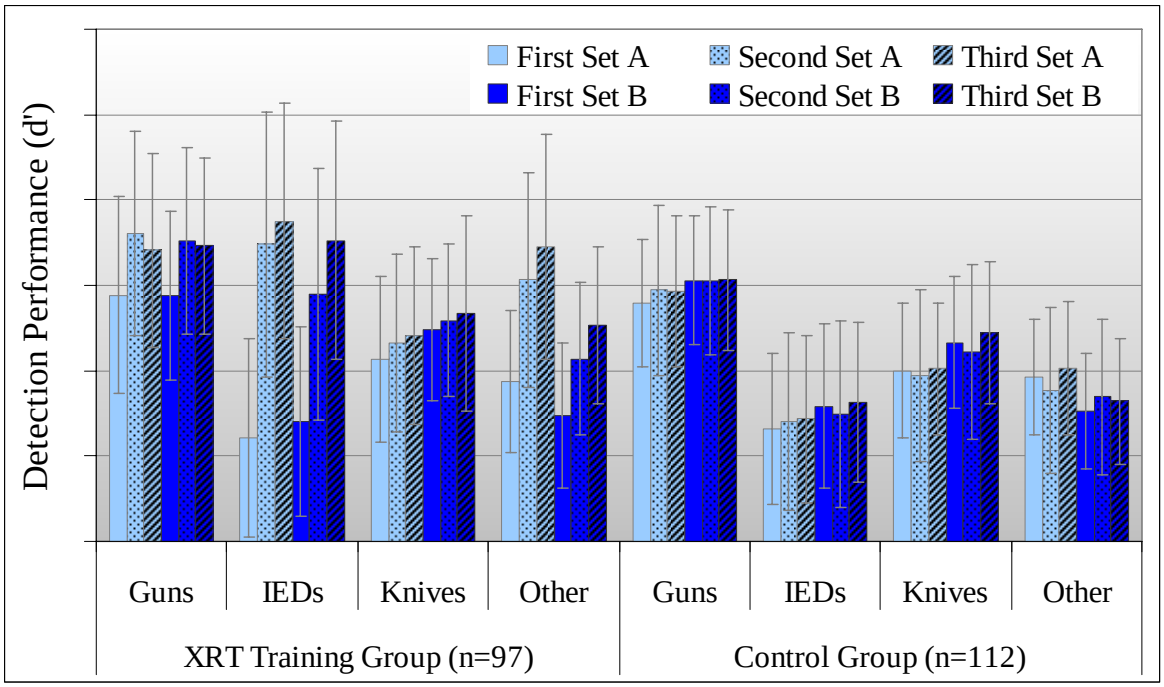

Figure 6.8. Detection performance with standard deviations for the XRT training group vs. the control group comparing first, second and third measurement for set A and set B and each threat category separately.

The extended ANOVA with the additional within-participant factor threat category revealed the main effects and interactions as specified in Table 4.3c. The main effect of set was not significant but there were significant interactions with set (see Table 4.3c). However, as can be seen in Figure 6.8, these interactions are rather small, which implies large transfer effects.

Figure 6.9 shows the results of the viewpoint analysis. An ANOVA was conducted on d' scores with the within-participant factors measurement, threat category and viewpoint and the between-participant factor group. It showed significant main effects of measurement, category, viewpoint and group. For details and interactions see table 4.3d. The large main effect of viewpoint indicates a higher detection performance for objects in easy (canonical) viewpoint compared to objects presented in a difficult (rotated) view (cf. Fig. 9). However, no significant interaction between viewpoint and training could be found. This would suggest that the viewpoint effect is unaffected by the training and could not be decreased. Pairwise *t*-tests showed a

significant increase in detection performance at the second measurement for both views in all categories for the XRT training group with the exception of knives in the easy view (p = .53, d = 0.07). All other comparisons were significant p < .05, d > 0.31). For the control group no significant increase in detection performance could be found (all p > .10, d < .0.19), see Table 4.5 for details. Training with XRT has an effect not only on the objects in the easy view but also on those in the difficult view. The screeners could make the association between the rotated object they detected during training and the canonical view of the object which is displayed in the object information in XRT.

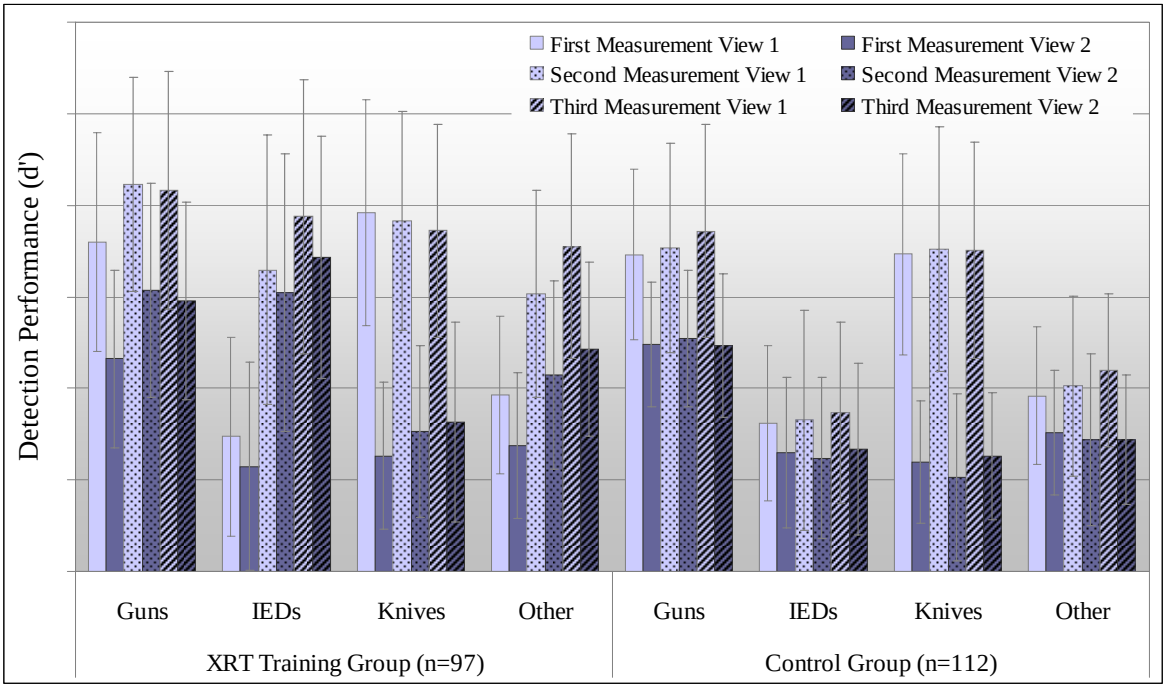

Figure 6.9. Detection performance with standard deviations for the XRT training group vs. the control group comparing first, second and third measurement for both views and each threat category separately.

Table 4.5

Results of the t-tests comparing the detection performance of the four categories for easy view (V1) and difficult view (V2) between the first (t1) and second (t2) measurement

XRT training group	***t*(96)**	***p***	***d***
Guns: V1t1 – V1t2	-4.21	< .01	0.53
IEDs: V1t1 – V1t2	-12.25	< .001	1.42
Knives: V1t1 – V1t2	0.64	= .53	0.07
Other: V1t1 – V1t2	-8.95	< .001	1.12
XRT training group	***t*(96)**	***p***	***d***
Guns: V2t1 – V2t2	-6.03	< .001	0.70
IEDs: V2t1 – V2t2	-11.45	< .001	1.43
Knives: V2t1 – V2t2	-2.53	< .05	0.31
Other: V2t1 – V2t2	-6.17	< .001	0.84
Control group	***t*(111)**	***p***	***d***
Guns: V1t1 – V1t2	-0.21	= .84	0.02
IEDs: V1t1 – V1t2	-0.76	= .45	0.08
Knives: V1t1 – V1t2	-0.66	= .51	0.07
Other: V1t1 – V1t2	-1.26	= .21	0.13
Guns: V2t1 – V2t2	-0.67	= .50	0.09
IEDs: V2t1 – V2t2	0.71	= .48	0.07
Knives: V2t1 – V2t2	1.65	= .10	0.19
Other: V2t1 – V2t2	0.64	= .53	0.07

In summary, a large and significant training effect was found for the group who trained with XRT for three and six months compared to the control group who used another CBT for the same time. A significant training effect has been observed for all four threat categories (guns, knives, IEDs and other), whereas the extent of the effect varied between categories. A large transfer of the acquired knowledge about the visual appearance of trained objects (set A) to untrained but similar looking objects (set B) was found for the XRT training group but not for the control group. This means that training with XRT helped screeners to detect other prohibited items which were not part of the training. Substantial effects of viewpoint effect could be

observed, i.e. unusual views of prohibited objects were much harder to detect than canonical views.

8.4 EXPERIMENT 2

The main aim of Experiment 2 was to replicate the results of Experiment 1 at another European airport. In addition, another conventional CBT was used for the control group. Thus it could be investigated whether conventional CBTs differ from each other regarding training effectiveness compared to XRT.

8.4.1 Method

8.4.1.1 Participants

163 airport security screeners of another mid-size European airport participated in Experiment 2. All screeners conducted the X-Ray CAT 1.0.0 three times with an interval of three months between the measurements. The adaptive CBT group (XRT group) consisted of 84 screeners who conducted weekly recurrent CBT using X-Ray Tutor (XRT) CBS 2.0 Standard Edition between all three test measurements. The control group consisted of 79 screeners and they used another conventional CBT than the control group of Experiment 1. As in Experiment 1, according to the security organization and their Appropriate Authority, airport security screeners of both groups conducted about 20 min CBT per week. Analysis of XRT training use showed that on average, each screener trained 20.92 minutes (*SD* = 2.87) per week.

8.4.1.2 Material and Procedure

Materials and procedure in Experiment 2 were the same as in Experiment 1. Again, all screeners took the X-Ray CAT at the beginning and after three and six months of CBT. The only difference was the conventional CBT for the control group, which was another one than in Experiment 1. As well this CBT is used at many airports worldwide. As the conventional CBT used in Experiment 1, this CBT has a much smaller threat image library than XRT and threat objects are not displayed in many different view. As well this CBT includes no individually adaptive trainings algorithm and there is a fixed combination of threat items in bags.

8.4.2 Results and Discussion

This section is structured the same way as in Experiment 1. Figure 6.10 shows the detection performance d’ for both groups and all three test measurements. As in

Experiment 1, individual d´ scores were subjected to repeated measures ANOVA with the within-participant factor measurement (first, second and third) and the between-participant factor group (XRT training group and control group). Again, there were large main effects of measurement $\eta^2 = .50$, $F(2, 322) = 163.52$, $p < .001$, group, $\eta^2 = .26$, $F(1, 161) = 56.34$, $p < .001$, and a significant interaction of measurement and group $\eta^2 = .33$, $F(2, 322) = 78.40$, $p < .001$. The large interaction is consistent with Figure 6.10 showing a much larger performance increase as a result of training for the XRT training group when compared to the control group. This was confirmed by independent samples *t*-tests. There was no significant difference between both groups for the first measurement $t(161) = -.22$, $p = .83$, $d = 0.03$, but a highly significant difference already in the second measurement $t(161) = 6.66$, $p < .001$, $d = 1.05$ after three months of training. As in Experiment 1, additional paired-samples *t*-tests revealed significant differences for the XRT training group between all measurements. In contrast to Experiment 1, there were also significant differences for the control group between the first and second measurement, although not between the second and third measurement (see Table 4.6). Thus, the conventional CBT used in Experiment 2 did also result in increased detection performance although substantially less than XRT.

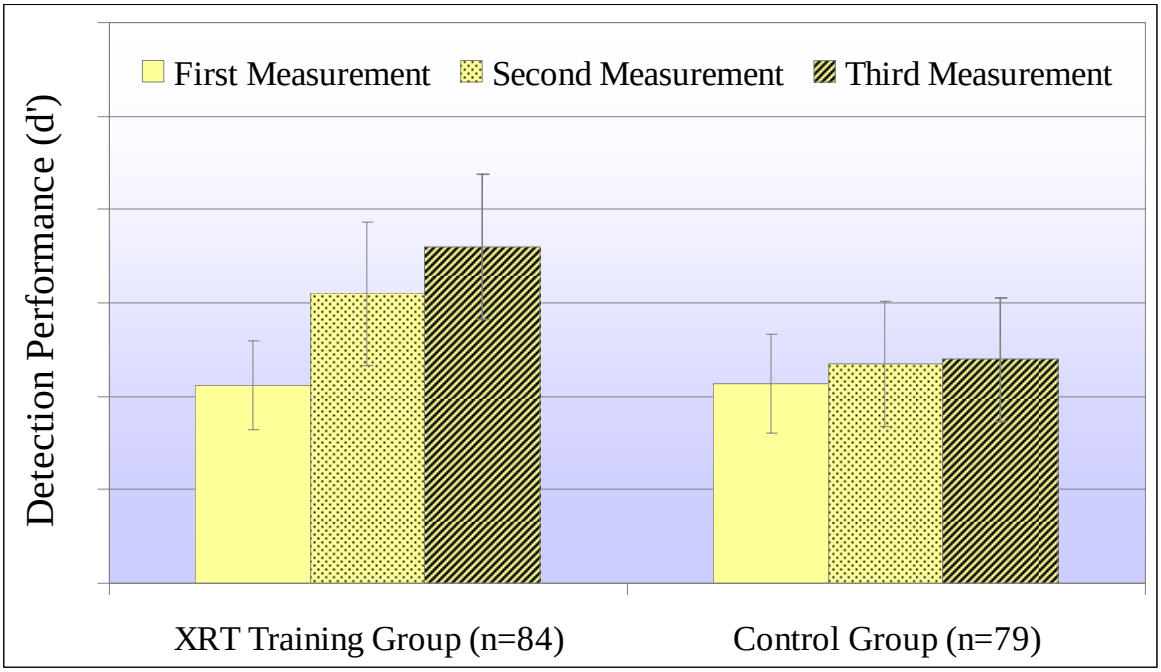

Figure 6.10. Detection performance with standard deviations for the XRT training group vs. the control group comparing first, second and third measurement.

Table 4.6

Results of the t-tests comparing the detection performance of first (t1), second (t2) and third (t3) measurement

	***t*(83)**	***p***	***d***
XRT Training Group (t1 – t2)	-12.21	< .001	1.57
XRT Training Group (t2 – t3)	-7.07	< .001	0.65
	***t*(78)**	***p***	***d***
Control Group (t1 – t2)	-3.67	< .001	0.36
Control Group (t2 – t3)	-0.91	= .37	0.07

Figure 6.11 shows the detection performance of both screener groups broken up by prohibited item category and the three test measurements. Again, a clear effect of training on the detection performance can be seen for the XRT training group with the largest increase after the first three months of training. However, also the control group shows a slight increase in detection performance at least for the second measurement. The analysis of variance (ANOVA) with threat category as additional within-participant factor showed significant main effects and significant interactions (for details see Table 4.7a). The results are comparable to those in Experiment 1.

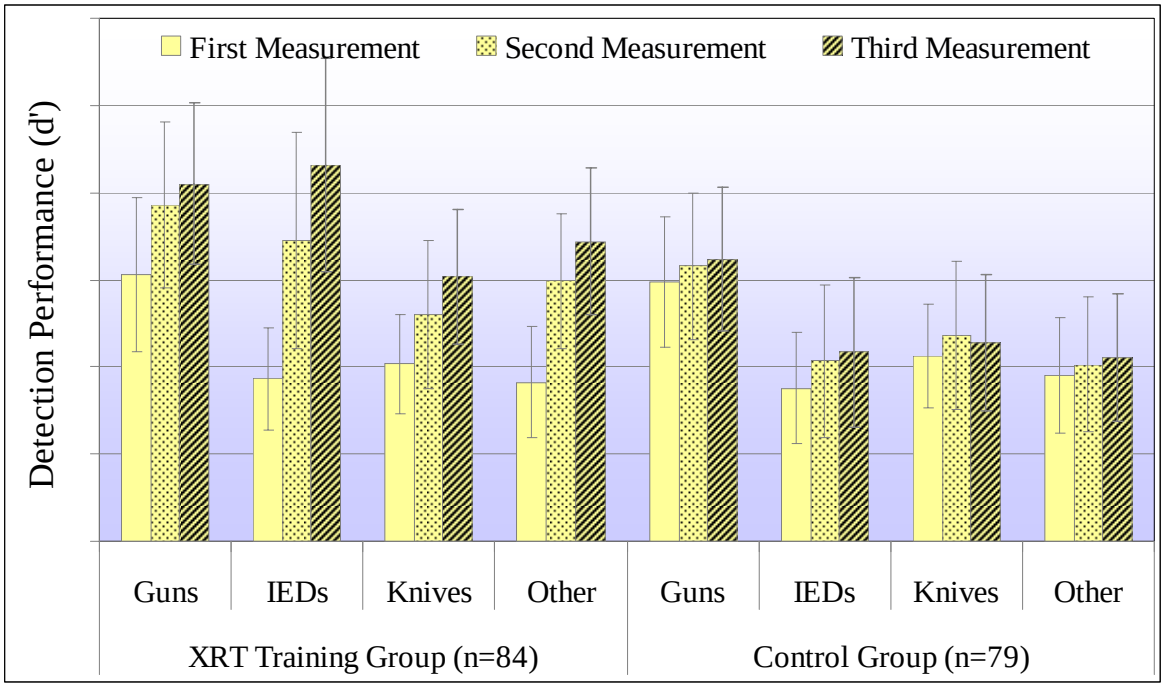

Figure 6.11. Detection performance with standard deviations for the XRT training group vs. the control group comparing first, second and third measurement for each threat category separately.

Most importantly, detection of guns was best initially, while detection of IEDs was much lower. After six months of recurrent adaptive CBT, screeners of the XRT training group could detect IEDs even slightly better than guns. This nice replication of the results obtained in Experiment 1 clearly shows that IED detection is not difficult per se but only a matter of the right training. As mentioned above this conclusion can only be made for multi-component IEDs. As in Experiment 1 detection of guns was best, followed by knives. Both threat categories contain threat objects which are relatively often seen at checkpoints. Again job experience, that is comparing X-ray images and the content of passenger bags, seems to have a considerably influence on the detection performance. Nevertheless, training improves the performance significantly which shows that job experience alone is not enough to store all kind of shapes. As shown in Table 4.8, *t*-tests between the first and second measurement revealed significant training effects for the XRT training group for all threat categories with large effect sizes (all $d > .0.80$). In contrast to Experiment 1, there were also significant effects for the control group, although with rather low effect sizes (all $d < 0.56$). Thus the conventional CBT used in Experiment 2 also resulted in performance increases although much less than XRT.

Table 4.7

Results of the ANOVAs in Experiment 2

	Factor	**df**	**F**	**η^2**	***p***
	Measurement (M)	2, 322	160.78	.50	< .001
	Threat Category (T)	3, 483	234.85	.59	< .001
	Group (G)	1, 161	64.98	.29	< .001
a)	M x G	2, 322	78.54	.33	< .001
	T x G	3, 483	37.63	.19	< .001
	M x T	6, 966	26.24	.14	< .001
	M x T x G	6, 966	16.67	.09	< .001
	Measurement (M)	2, 322	156.12	.49	< .001
	Set (S)	1, 161	58.45	.27	< .001
	Group (G)	1, 161	56.03	.26	< .001
b)	M x G	2,322	82.16	.34	< .001
	M x S	2, 322	8.88	.05	< .001
	S x G	1, 161	31.37	.16	< .001
	M x S x G	2, 322	15.52	.09	< .001

	Factor	df	F	η^2	*p*
	Measurement (M)	2, 322	162.28	.50	< .001
	Set (S)	1, 161	41.88	.21	< .001
	Threat Category (T)	3, 483	231.83	.59	< .001
	Group (G)	1, 161	71.93	.31	< .001
	M x G	2, 322	84.18	.34	< .001
	M x T	6, 966	27.50	.15	< .001
	M x S	2, 322	11.42	.07	< .001
c)	S x G	1, 161	36.23	.18	< .001
	S x T	3, 483	33.59	.17	< .001
	T x G	3, 483	40.15	.20	< .001
	M x T x G	6, 966	16.87	.10	< .001
	M x S x G	2, 322	10.09	.06	< .001
	M x S x T	6, 966	1.48	.01	= .18
	S x T x G	3, 483	3.69	.02	< .05
	M x S x T x G	6, 966	2.64	.02	< .05
	Measurement (M)	2, 322	152.62	.49	< .001
	View (V)	1, 161	1849.85	.92	< .001
	Threat Category (T)	3, 483	216.74	.57	< .001
	Group (G)	1, 161	70.32	.30	< .001
	M x G	2, 322	80.05	.33	< .001
	M x T	6, 966	26.57	.14	< .001
	M x V	2, 322	2.99	.02	= .05
d)	V x G	1, 161	0.62	.00	= .43
	V x T	3, 483	288.98	.64	< .001
	T x G	3, 483	34.91	.18	< .001
	M x T x G	6, 966	14.95	.09	< .001
	M x V x G	2, 322	1.21	.01	= .30
	M x V x T	6, 966	2.82	.02	< .05
	V x T x G	3, 483	1.69	.01	= .17
	M x V x T x G	6, 966	1.89	.01	= .08

Table 4.8

Results of the t-tests comparing the categories between first (t1), second (t2) and third (t3) measurement

XRT training group	***t***	**df**	***p***	***d***
Guns t1 – t2	-6.01	83	< .001	0.86
IEDs t1 – t2	-12.84	83	< .001	1.74
Knives t1 – t2	-5.81	83	< .001	0.80
Other t1 – t2	-12.30	83	< .001	1.64
Guns t1 – t3	-8.19	83	< .001	1.15
IEDs t1 – t3	-20.22	83	< .001	2.70
Knives t1 – t3	-10.97	83	< .001	1.48
Other t1 – t3	-16.46	83	< .001	2.18
Control group	***t***	**df**	***p***	***d***
Guns t1 – t2	-2.19	78	< .05	0.23
IEDs t1 – t2	-3.60	78	< .01	0.42
Knives t1 – t2	-2.73	78	< .01	0.33
Other t1 – t2	-1.46	78	< .15	0.18
Guns t1 – t3	-2.72	78	< .01	0.34
IEDs t1 – t3	-4.61	78	< .001	0.56
Knives t1 – t3	-2.05	78	< .05	0.23
Other t1 – t3	-2.59	78	< .05	0.30

By an ANOVA with measurement and set as within-participant factors and group as between-participant factor, we investigated if training effects can also be shown for threat objects which were not included in the training sessions. There were main effects and interactions for all factors showing similar results as in Experiment 1 (see Table 4.7b for details). As in Experiment 1, a large transfer effect was found (see Figure 6.12). Not only for the prohibited items of set A, which were included in the training library of XRT, but also for the untrained prohibited objects of set B, screeners of the XRT training group showed a large increase in detection performance after training. Paired-samples *t*-tests between the first and second measurement showed training effects for both sets and also for both groups whereas again large effect sizes were found for the XRT training group and small effect sizes

for the control group (trained group set A: $t(83) = -13.10$, $p < .001$, $d = 1.77$ and set B: $t(83) = -9.53$, $p < .001$, $d = 1.24$, control group set A: $t(78) = -2.32$, $p < .05$, $d = 0.24$ and set B: $t(78) = -3.00$, $p < .01$, $d = 0.32$). Pairwise *t*-tests showed no significant difference in the difficulty of set A and Set B for both groups at the first measurement (XRT training group: $t(83) = 1.16$, $p = .25$, $d = 0.10$, control group: $t(78) = 1.93$, $p = .06$, $d = 0.19$).

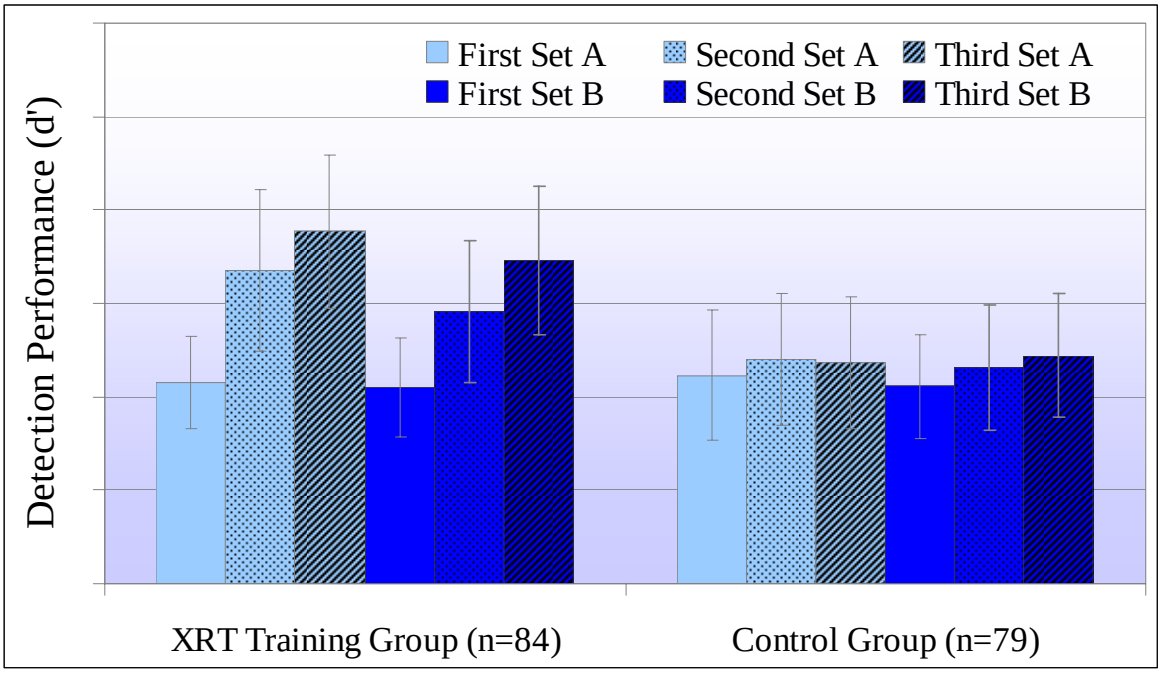

Figure 6.12. Detection performance with standard deviations for the XRT training group vs. the control group comparing first, second and third measurement for set A and set B separately.

Figure 6.13 includes also the threat category in the analysis. Paired samples *t*-tests were calculated in order to investigate if the training effect between the first and second measurement was significant for each category in both sets for the XRT training group. Results revealed significant effects for all categories in each set ($p < .01$, $d = 0.51$ for knives in Set B, $p < .001$, $d > 0.74$ for all other categories). Thus, as in Experiment 1, XRT resulted in large detection performance increases even for prohibited objects that are not part of the XRT image library (X-Ray CAT image set B). For the control group the difference between the first and *third* measurement was calculated in order to maximize the chances for finding a significant training effect. The following *t*-tests were significant: IEDs for both sets, knives only for set A, and other threat objects for both sets ($p < .05$, $d > 0.23$). All other values were not significant ($p > .06$, $d < 0.28$) and reveal no effect of training between the different measurements.

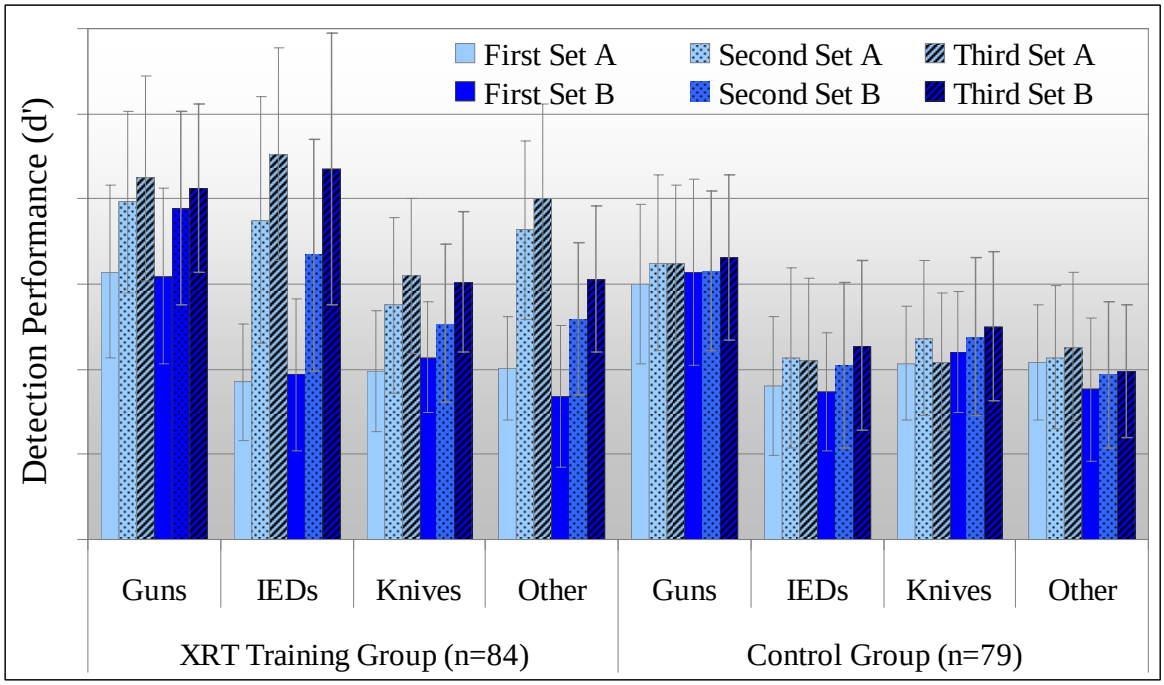

Figure 6.13. Detection performance with standard deviations for the XRT training group vs. the control group comparing first, second and third measurement for set A and set B and each threat category separately.

As in Experiment 1, individual d´ scores were subjected to an extended ANOVA with the within-participant factors measurement, X-Ray CAT image set, threat category and the between-participants factor group. All main effects and interactions were significant except the interaction between measurement, set and threat category (see Table 4.7c for details). In contrast to Experiment 1 the ANOVA revealed a main effect of set and significant interactions with set. However, as can be seen in Figure 6.13 they were rather small, which implies large transfer effects. As in Experiment 1 the results clearly show a training effect for each category and in both sets. This is consistent with the results of the *t*-tests explained above. The training effect that was found for the control group revealed itself also in the sets, that is, there was a transfer effect for the control group, too.

Last, the effect of viewpoint was investigated calculating a four-way ANOVA. Results show clear main effects of measurement, view, threat category and group. For details on interactions please refer to Table 4.7d. As illustrated in Figure 6.14, the detection performance is clearly much higher for objects that are shown in the easy view (View 1) than for the objects that are shown from an unusual viewpoint (View 2). This effect is valid for all threat categories and for the XRT training group as well as for the control group. However, the viewpoint effect is not the same for different

threat categories. The graphs in Figure 6.14 suggest that the largest viewpoint effect can be observed for the detection of knives, the smallest one for IEDs.

As in Experiment 1, pairwise *t*-tests showed a significant increase in detection performance at the second measurement for both views for the XRT training group for all four threat categories ($p < .01$, $d > 0.49$. For the easy view, the control group showed a significant effect for IEDs only ($p < .05$, $d = 0.32$), all other *t*-tests were not significant ($p > .07$, $d < 0.25$). For the difficult view all *t*-test with one exception were significant for the control group ($p < .05$, $d > 0.26$). Only the training effect of knives in the rotated view was not significant $p = .07$, $d = 0.24$ (see Table 4.9 for details). But the results show that although some significant effects in the control group were observed, effect sizes were small compared to those of the XRT training group.

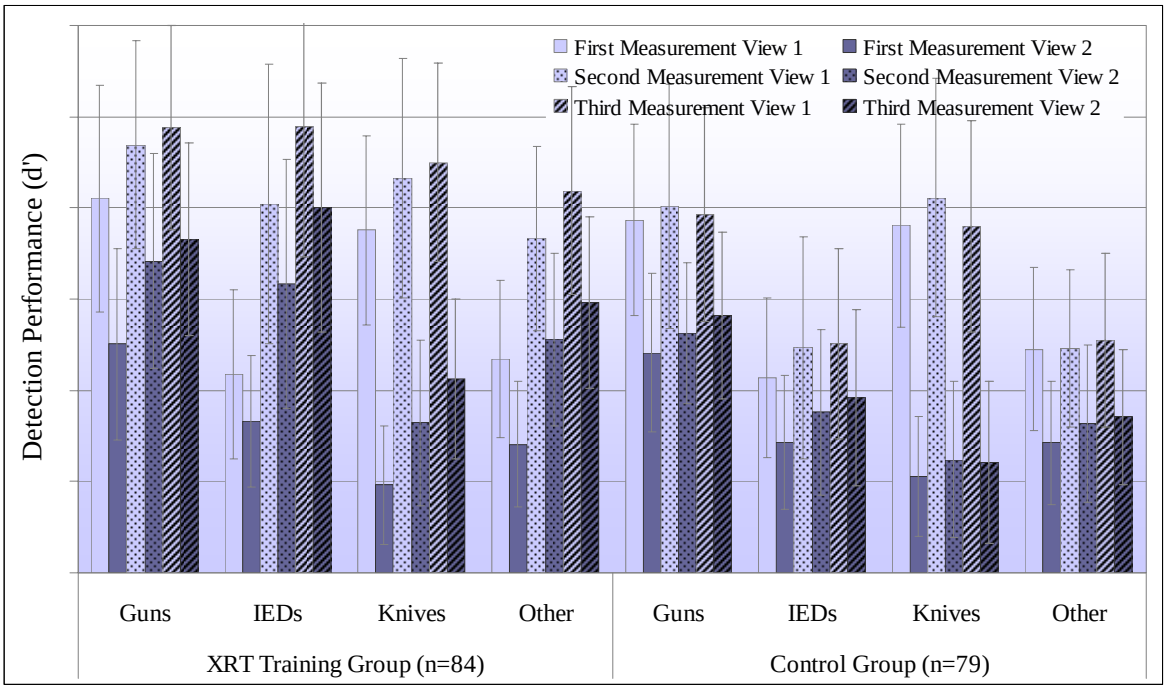

Figure 6.14. Detection performance with standard deviations for the XRT training group vs. the control group comparing first, second and third measurement for both views and each threat category separately.

Table 4.9

Results of the t-tests comparing the detection performance of the four categories for easy view (V1) and difficult view (V2) between the first (t1) and second (t2) measurement

XRT training group	***t*(83)**	***p***	***d***
Guns: V1t1 – V1t2	-3.59	< .01	0.49
IEDs: V1t1 – V1t2	-10.93	< .001	1.51
Knives: V1t1 – V1t2	-4.35	< .001	0.48
Other: V1t1 – V1t2	-9.79	< .001	1.42
Guns: V2t1 – V2t2	-5.46	< .001	0.82
IEDs: V2t1 – V2t2	-9.99	< .001	1.45
Knives: V2t1 – V2t2	-5.79	< .001	0.88
Other: V2t1 – V2t2	-10.33	< .001	1.40
Control group	***t*(78)**	***p***	***d***
Guns: V1t1 – V1t2	-1.07	= .29	0.13
IEDs: V1t1 – V1t2	-2.64	< .05	0.32
Knives: V1t1 – V1t2	-1.87	= .07	0.25
Other: V1t1 – V1t2	-0.05	=.96	0.01
Guns: V2t1 – V2t2	-2.35	< .05	0.26
IEDs: V2t1 – V2t2	-3.24	< .01	0.41
Knives: V2t1 – V2t2	-1.81	= .07	0.24
Other: V2t1 – V2t2	-2.11	< .05	0.28

In summary, very similar results as in Experiment 1 have been found in Experiment 2. A large and significant training effect was observed for the group who trained with XRT compared to a control group who used a conventional CBT for the same time. A significant training effect has been observed for all four categories (guns, knives, IEDs and other) for the XRT training group, whereas the effect size varied between categories. Also a large transfer of the acquired knowledge about the visual appearance of trained objects (set A) to untrained but similar looking objects (set B) was found for the XRT training group. Additionally a viewpoint effect could be observed which shows that unusual views of forbidden objects are much harder to detect than canonical views. In contrast to Experiment 1, the control group also showed increases of detection performance, which implies that the conventional CBT

used in Experiment 2 is more effective than the one used in Experiment 1. Moreover, there was also a transfer effect for the control group.

8.5 GENERAL DISCUSSION

The first aim of this study was to investigate how well airport security screeners can detect guns, knives, IEDs and other prohibited items in X-ray images of passenger bags. Two experiments conducted at two European airports provided very similar results. A computer-based test (X-Ray CAT) was conducted before and after three and six months of weekly (about 20 min per screener) CBT at each airport. The first measurement revealed that guns were detected best, followed by knives, other prohibited items and IEDs. This result implies that job experience has already an effect on the detection of threat items. Threat objects like guns and knives which are relatively often taken along by passengers and therefore more often seen at checkpoints are detected better than IEDs which are normally not seen on the job. Nevertheless, the effect of job experience is quite small compared to a trainings effect which can be achieved with an individually adaptive CBT. In both experiments and airports, one group used an adaptive CBT (XRT) with individually adaptive algorithms, a large library of prohibited items depicted in a variety of different views, and automatically created prohibited item to bag combinations. The other group used a conventional CBT system with no adaptive algorithms, a smaller image library, and fixed combinations of threat items in bags. While XRT was used in both experiments and airports, two different conventional CBT systems were used for the control groups of Experiment 1 (airport 1) and Experiment 2 (airport 2). At both airports, XRT training group results revealed a training effect for all types of threat objects (guns, knives, IEDs, and other prohibited items). However, effect sizes differed remarkably for the four categories. While guns were detected best and IEDs were detected worst at the beginning, IED detection of the XRT training group was as good, or even slightly better, than gun detection after several months of training. This shows that the detection of IEDs is not difficult per se, but rather depending on the training of screeners. However, all IEDs used in this study contained a detonator, wires, explosive, a triggering device and a power source. Therefore, these conclusions are only applicable to the detection of such multi-component IEDs. A large training effect for IEDs can be expected because they are usually not encountered at airport security checkpoints and therefore not known to screeners

without enhanced training in IED detection. In a study with hold baggage screeners, large training effects for IEDs were also found, which is very consistent with results of this study (Schwaninger & Hofer, 2004). The relatively large training effect for the category "other" which includes self defense gas spray, electric shock devices etc. might be also explained by less on the job exposure of these prohibited items. In contrast to IEDs and other prohibited items, guns seem to be well known by screeners either because of their typical shape or the frequency by which they are encountered at the airport security screening checkpoint (e.g. toy guns). Therefore, detection performance before training is already high for guns and a large improvement is impossible. It is also noticeable that detection for knives showed the smallest training effect in both experiments. Although the detection was at the baseline measurement higher than for IEDs and other prohibited items, after six months of training screeners' performance was poorest for knives. On average, knives are smaller than IEDs and other threat items and show less diagnostic features. This might be a reason for the lower detection performance increase for this threat category.

While training with XRT resulted in large training effects, the tested conventional CBT systems were less effective. In Experiment 1, there were no training effects at all, while only small training effects were observed for the conventional CBT system used in Experiment 2. This could be due to one or a combination of the following reasons: First, the conventional CBT systems tested in this study do not feature individually adaptive training algorithms like XRT. Second, in contrast to XRT, the conventional CBT systems did not contain such a large image library with many prohibited items depicted from a variety of different viewpoints, especially IEDs. Third, while in XRT prohibited items are blended into X-ray images of passenger bags on the fly using scientifically validated and individually adaptive algorithms based on image measurement as described in Schwaninger et al. (2007), the conventional CBT systems used in Experiment 1 and 2 have only fixed combinations of prohibited items in bags. Finally, we had to rely on the statement of the appropriate authority and the security companies regarding the amount of training that was conducted by screeners of the control group and the XRT training group, which should have been on average 20 min per week per screener. Analysis of XRT training data showed, that this was clearly fulfilled for screeners of the XRT training group at both airports.

Since the X-Ray CAT is composed of two comparable (similar looking) sets (set A and set B) whereof only the threat objects of set A were included into the XRT training

system, transfer effects can be tested, i.e. whether training with certain prohibited items helps increasing detection of other prohibited items that are not contained in the training. Overall, the comparison of the two sets A and B at the baseline measurement (before training) shows no significant difference. However, in Experiment 1 there was a slight difference for the control group between the two sets indicating that the two sets are not exactly equal in terms of image difficulty for this sample. But this possible objection to the transfer effect can be disapproved with two arguments: first, the effect size was only small according to the conventions by Cohen (1988) and second, only one of the two control groups showed a significant difference. Therefore, the transfer effect in the results of the XRT training group can be attributed to the training of set A only. The small training effect for the control group in Experiment 2 is also reflected in the detection increase of both sets after training. Although the conventional CBT system of this control group did not contain any objects from the test, the training with this training system apparently also leds to a transfer of the knowledge to the objects in the test. In another study it would be interesting to compare the objects that are comprised in the two training systems used by the control groups regarding their similarity to the test objects. Contrary to our results, Smith, Redford, Gent and Washburn (2005) found a large decrease in screeners' detection performance when specific trained objects were replaced with new images belonging to the same categories (see also Smith, Redford, Washburn, & Taglialatela, 2005). According to these authors, improvement in screening performance is attributable only to specific-token familiarity that developed for the original images and not to a category generalization. They state constraints on categorization and the use of category-general information when humans face visual complexity and have to identify targets within it. Our results can be interpreted in support of generalization of visual learning in X-ray image interpretation. However, it might be possible that the objects of the untrained set in our study are so similar to the trained objects that a specific-token familiarity led to the detection performance increase and not a true generalization effect. The lacking transfer effect in knives would along these lines mean that the objects in set A and set B are not similar enough in shape to generate a specific-token familiarity. Therefore only the learnt objects could generate a training effect but not the unlearnt ones. For Schwaninger and Hofer's (2004) findings of a large increase in detection performance of IEDs after recurrent CBT with other members of the category than those included in the test, it

would mean, that those objects were very similar in order to create a specific-token familiarity and therefore a training effect.

In both Experiments a large viewpoint effect was also revealed. This is consistent with view-based theories of object recognition (for reviews see for example Tarr & Bülthoff, 1995a, 1998; Graf et al., 2002; Hayward, 2003). After training, easy and difficult views were recognized much better. Interestingly, there was no significant interaction between measurement and viewpoint, i.e. although training resulted in improved performance for difficult views, the viewpoint effect (impairment for unusual vs. canonical views) remained stable even after six months of training. However, it must be pointed out that the XRT training algorithm only provides the screeners with unusual views of objects once a screener can detect a prohibited item well when depicted from easy perspective. That is, when screeners start to train with XRT all threat objects are shown in easy views. Only if these objects are detected reliably, the difficulty level is increased for a certain threat item by showing it in more difficult views (Schwaninger, 2004b). Thus, it is unclear whether a significant interaction between viewpoint and measurement would have been observed if the training duration would have been increased (e.g. to one year). The conclusion stands to reason that recognition of forbidden objects in X-ray images is dependent on exposure which has very important implications for an adaptive training system. It has been assumed that different views of each object become associated with one another during object rotation, either through active learning or through passive experiencing of the successive appearance of nearby views (Földiák, 1991; Stryker, 1991). Hence, it is important that during training screeners are getting feedback which forbidden object has been detected or missed. This feedback shows the photograph and also the X-ray image of that forbidden object always in the canonical view whereas the forbidden object merged into a bag is presented in different viewpoints. This leads to an association between an unusual view of an object and the canonical view which results in a sequential pairing of these views with each other (Wang, Obama, Yamashita, Sugihara, & Tanaka, 2005). This association, which forms during learning, is thought to underlie object recognition ability across changes in viewing angle (Palmeri & Gauthier, 2004).

For our future studies, it could also be interesting to increase the interval between the end of training and the testing of training transfer, as corresponding literature usually tests transfer of training after a considerable period of time in order to measure the stability of the transfer (e.g., Saks & Belcourt, 2006). In any case, our

findings show that the knowledge about the visual appearance of forbidden objects, which airport security screeners acquire during recurrent CBT, can be transferred to similar looking, but not previously seen objects and also the effect that rotated views are much harder to detect can be decrease with training. To make sure that objects are well detected it is important that a large and representative image library of prohibited objects is used and that these objects are learned from different viewpoints. Additionally the library should be updated constantly to adapt to new threats. Overall, this study has shown that adaptive CBT can be a powerful tool to increase screeners' X-ray image interpretation competency in an efficient and effective way.

PART III

AGE-EFFECTS IN AVIATION SECURITY SCREENING

9. USE IT AND STILL LOSE IT: THE INFLUENCE OF AGE AND JOB EXPERIENCE ON DETECTION PERFORMANCE IN X-RAY SCREENING

9.1 ABSTRACT

In recent years, research on cognitive aging increasingly focuses on cognitive development across middle adulthood. However, still little is known about the longterm effects of intensive job-specific training of fluid intellectual abilities. We examined the effects of age and job-specific practice of cognitive abilities on detection performance in airport X-ray security screening. In experiment 1 (N = 308; 24-65 years), we examined performance in the X-Ray Object Recognition Test, a speeded visual detection task in which participants have to find threat items in X-ray images of passenger bags. In experiment 2 (N = 155; 20-61 years) object recognition that was closer to the practical task of baggage screening was evaluated. Results from both experiments show high performance in older adults and significant negative age correlations that cannot be overcome by more years of job-specific training. We discuss the implications of our findings for theories of lifespan cognitive development and training concepts[21].

9.2 INTRODUCTION

In recent years, cognitive aging research is increasing its focus on the effects of midlife development on developmental changes in old age. Part of the renewed interest in midlife development were longitudinal results from representative population samples suggesting cognitive performance to be differentially variable and showing little covariation between changes across domains of cognitive functioning (Martin & Zimprich, 2005). However, whereas changes in experience-related knowledge seem to be small across middle age, fluid abilities demonstrate average declines from the ages of 30 onward (e.g., Schaie, 2005). These average declines might be due to the variability in job-induced stimulation of fluid abilities across individuals within representative population samples and the typically low

[21] I gratefully acknowledge the help of Mike Martin and Adrian Schwaninger in preparing the manuscript. I thank Judith Riegelnig for providing the data of this study.

experimental control over job-related cognitive stimulation in representative samples. Therefore, an examination of the effects of job-related stimulation of fluid abilities in aviation security screening was examined.

X-ray screening of passenger bags is a highly demanding task which includes both, experience related knowledge and fluid abilities. Schwaninger et al. (2005) defined knowledge-based and image-based factors to be important in the X-ray screening task. Screeners need to know which items are prohibited and what they look like in X-ray images. Such knowledge-based factors are related to the memory component of visual object recognition and depend strongly on training (Schwaninger & Hofer, 2004). Image-based factors include the effects of superposition and viewpoint of the threat item, as well as bag complexity (Schwaninger et al. 2005; see Figure 1.3). Threat objects are more difficult to recognize when superimposed by other objects (effect of superposition) or when depicted from a difficult viewpoint (effect of viewpoint). Furthermore, it is more difficult to detect a threat item in a close-packed bag as other objects in the bag distract attention (effect of bag complexity). These factors are more related to abilities of visual and spatial cognition. Thus, processes such as visual search, spatial imagination, working memory, attention and perception which are associated with an age-related decline should be important determinants in X-ray screening performance, and older adults should show lower performance compared to younger adults.

Previous findings on visual search and aging could show that older people perform slower and less accurate to locate targets than younger ones as the number of distractors increases (Humphrey & Kramer, 1997; Kramer & Atchley, 2000). In addition, older adults make more eye movements and tend to fixate areas for a longer time (Scialfa, Thomas, & Joffe, 1994). Laberge and Scialfa (2005) even assumed that the tendency to examine areas repeatedly reflect age-related declines in visual working memory. The influence of aging on spatial imagination or mental rotation tasks have been investigated in many research studies (e.g., Berg, Hertzog, & Hunt, 1982; Campos, Pérez-Fabello, & Gómez-Juncal, 2004; Dror & Kosslyn, 1994). Applying a spatio-visual capacity test, Campos et al. (2004) could show that imagining and conceiving objects in three dimensions is affected by aging. Likewise, Dror and Kosslyn (1994) found a relatively impaired image rotation of the elderly. Furthermore, they reported a decrease of image activation with aging, i.e., the process of accessing and activating visual memories. This image activation process can be supposed to be rather important in the X-ray image interpretation task. In

addition, spatial cognition tasks are often related to endogenous sex steroids, in particular testosterone and estradiol. Janowsky, Oviatt, and Orwoll (1994) reported better performance in a spatial cognition task when elderly people had received an androgen supplementation. Also Hampson (1995) reviewed some recent evidence suggesting that testosterone treatment is often associated with significant improvement in spatial cognition. As androgen level changes across the lifespan, age can impair spatial cognition. Note however, that evidence for a beneficial effect of higher androgen levels on cognition in older men could not be found by Wolf and Kirschbaum (2002). Overall, findings on age differences in abilities related to visual screening tasks suggest age differences with lower screening performance in old versus young adults.

In a demanding task such as X-ray screening, working memory and attention could play an important role. During rush hours at larger airports, the decision whether a bag is OK (contains no threat item) or NOT OK (contains a threat item) has to be made within 3-5 seconds. Working memory is described as the central executive that processes information at a conscious level (Baddeley, 1986). Various studies reported a decrease of working memory capacity with aging (Cherry & Park, 1993; Salthouse & Babcock, 1991). Furthermore, screeners have to be constantly vigilant which requires sustained attention across time. Results concerning the relationship between aging and sustained attention are not consistent (for an overview see Roger & Fisk, 2001). However, Deaton and Parasuraman (1993) found lower performance for vigilance tasks with a cognitive component, such as identification or decision about an item. There is also evidence that aging has a negative influence on perception as age-related changes of eye structures lead to decrease in sharpness and brightness of visual stimuli (Cabeza, 2001). Furthermore, deficits in color sensitivity due to the loss of photoreceptors were reported (Fozard & Gordon-Salant, 2001). The decrease of receptors also influences visual processing in the periphery. An eye tracking study revealed a repeating search in the same location of elderly people compared to younger ones (Scialfa et al., 1994).

Although aging seems to affect cognitive processes, there is some evidence that older adults might be able to compensate for their cognitive deficits using their working experience adopting more efficient strategies (Zec, 1995). In addition, one may assume that extensive practice with a task as is the case with experienced screeners having trained the task on the job for hundreds and thousands of hours

over many years should improve performance in all age groups (e.g., Kliegl, Smith, & Baltes, 1989; Maguire, Gadian, & Johnsrude, 2000).

In this study the effects of job-related stimulation of fluid abilities in airport security screeners exposed to extensive amounts of speeded visual search and detection tasks was examined. This was designed to answer four questions: First, are critical levels of detection performance achieved across an age range from 20-65 years of age. We expect that despite potential age effects, all age groups reach a high and critical level of performance. Second, do age-related differences in job-specific fluid performance exist when persons are practiced in that ability and may use any strategy available in the workplace to maximize their performance. Based on findings of age-related declines in fluid abilities across middle age, age differences in screening performance are expected. However, if on-the-job practice of the required skills may compensate for these declines, then no age effects are expected. Third, do older workers profit more from job experience than younger workers. It may be the case that practicing screening skills is more important for screening performance in old versus young age, because of the importance of practice to overcome deficits in underlying abilities that are at peak in younger ages. Fourth, one may argue that in reality threat items occur rather seldom and that, therefore, larger amounts of practical experience of the older screener could not be used. Therefore, it was investigated if age effects differ depending on the task demands, i.e., when a frequent decision about the threat potential of an item versus the rare detection of a threat event are required.

9.3 EXPERIMENT 1

Experiment 1 examined whether fluid abilities which are needed in X-ray screening tasks decline across an age range form 20 to 65 years and whether detection performance of experienced aviation security screeners is on a higher level than performance of novices. The influence of age on detection performance in an X-ray screening test as a speeded visual search and detection task was investigated. Additionally, a correlation should show if experience and practice might compensate a negative effect of aging and whether detection performance increases with more working experience.

9.3.1 Method

9.3.1.1 Participants

In Experiment 1, a total of 308 cabin baggage screening (CBS) screeners between 24 and 65 years of age (*M* = 50.28, *SD* = 9.43) working more than an average of 30 percent for at least 2.6 years (*M* = 10.38, *SD* = 5.56) participated in this study. Screeners who did not differ in job-related knowledge performed the X-Ray Object Recognition Test (X-Ray ORT), a speeded visual task in which participants have to decide for 256 images if it contains a threat item or not. An outlier analysis was performed and values higher or lower than two standard deviations from the mean were excluded, that is 17 screeners.

9.3.1.2 Materials and Procedure

For Experiment 1, a computer-based X-ray screening test, the X-Ray ORT was used (Hardmeier et al., 2005). The test was conducted in a well lit computer classroom with ten HP Compaq d530 CMT computers using the PCQuest software. The X-ray images were presented on full screen size at a resolution of 1024 x 768 pixels on a 17-inch TFT monitor. Brightness and contrast settings on each screen were set to 65 and 40, respectively.

The X-Ray ORT requires subjects to recognize guns and knives in passenger bags. It starts with a self-explanatory instruction including some exercise trials to familiarize the participants with the test taking procedure. After each image which is displayed for 4 seconds on the screen, screeners have to decide whether the bag is OK (no gun or knife in it) or NOT OK (a gun or knife in it). In addition, they have to indicate how sure they are in their decision clicking on a 90 point rating scale. The test takes about 45 minutes to complete. Varying bag complexity, superposition and the rotation of the threat item systematically, the X-Ray ORT measures visual abilities needed to cope with image-based factors in X-ray screening. Therefore, eight guns and eight knives are twice displayed in an easy and rotated view in bags with low and high complexity level. Half of the threat items are shown with little and half of them with high superposition. (for details see Hardmeier et al., 2005; Schwaninger et al., 2005). The test includes a total of 256 X-ray images in grayscale, half of them contain either a gun or a knife, and the other 128 images are harmless bags. The test is subdivided into four blocks, the order of blocks is counterbalanced across participants, and the order of trials within a block is random.

Reliability and validity measures of the X-Ray ORT were calculated based on 453 screeners (Cronbach Alpha measures > .89 and split half reliabilities > .78). Criterion-related validity was calculated correlating test results in the X-Ray ORT with Threat Image Projection (TIP) data. TIP data allows to measure on the job performance by displaying fictional threat items into real passenger bags (for more information about TIP data please see Experiment 2). Correlation of $r = .51$, $p < .01$ suggests that image-based factors measured by the X-Ray ORT are indeed important determinants of on the job performance in X-ray screening. For more details on reliability and validity see Hardmeier et al. (2005).

9.3.2 Results and Discussion

Detection performance in the X-Ray ORT was calculated using the detection performance measure d' (Green & Sweets, 1966). To avoid that a screener could reach a high hit rate by simply judging all bags as NOT OK d' takes into account the hit rate and the false alarm rate. According to the signal detection theory there are four possible outcomes depending on the presence of a threat item and the decision of the screener (see Figure 7.1). Judging a threat image as NOT OK results in a hit whereas judging a harmless bag as NOT OK results in a false alarm. Furthermore, judging a threat image as OK results in a miss and judging a harmless bag as OK results in a correct rejection. Hence, a good screener would detect nearly all threat items in passenger bags (high hit rate) and hardly ever send harmless bags to be hand-searched (low false alarm rate). D' is calculated by the formula z(hit) – z(false alarm). Furthermore, the detection performance measure d' is independent of the criterion. That is, if a screener is more anxious to miss an object, he judges more bags as NOT OK and thus both, the hit and false alarm rate increases.

	Threat item	No threat item
Decision bag NOT OK	***HIT***	***False Alarm***
Decision bag OK	***Miss***	***Correct Rejection***

Figure 7.1. Four possible answers when judging an x-ray image.

To make sure that critical levels of detection performance were achieved across an age range from 20-65 years of age, detection performance between experts and novices was compared (see Figure 7.2). Novices were 284 job applicants between 19 and 56 years (M = 38.61, SD = 10.14) who took the X-Ray ORT within the pre-employment assessment procedure. An analysis of variance (ANOVA) revealed a significant difference between these two groups η^2 = .28, $F(1, 590)$ = 230.24, $p < .01$. This result is also consistent with previous findings which showed that detection performance of aviation security screeners in the X-Ray ORT is on a higher level than those of novices (Schwaninger et al, 2005).

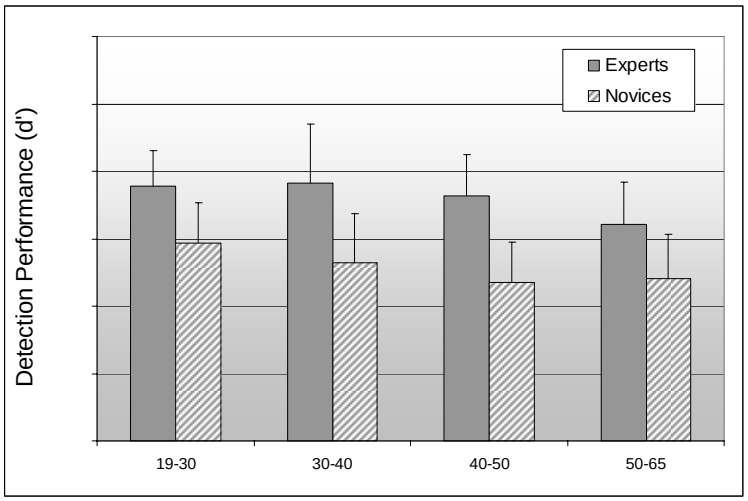

Figure 7.2. Detection performance d' of experts and novices in the X-Ray ORT for four age categories.

As expected, a partial correlation between detection performance d' and age of screeners controlling for years since employment showed a significant negative correlation between age and results in the X-Ray ORT (pr = -.27, $p < .01$). Figure 7.3 shows the correlation between d' and age in years. Results provide evidence that age-related differences in job-specific fluid performance even exist in when persons are practiced in that ability and may use any strategy available to maximize their performance. Thus on average, in the X-ray screening test, older screeners had a lower detection performance than younger ones which can be most probably related to several visual and cognitive functions which decline

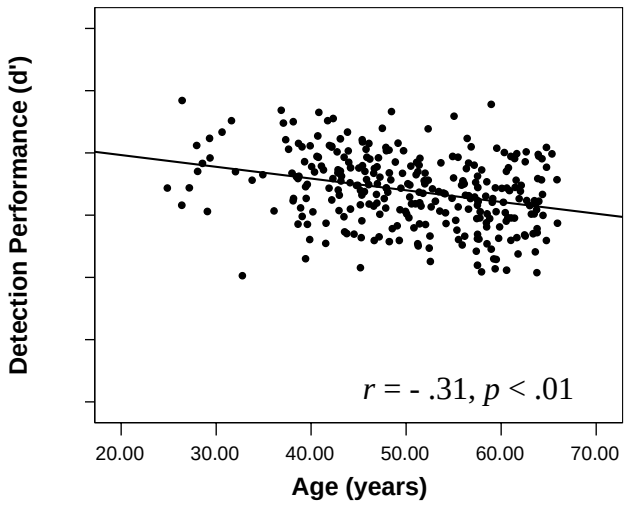

Figure 7.3. Correlation between detection performance (d') and age for cabin baggage screener.

with age. A partial correlation between d' in the X-Ray ORT and years since employment with age as control variable was also calculated. Result of pr = -.04, p = .48 indicates that detection performance in the X-Ray ORT can not be increased by on the job practice or experience and thus older screeners who are generally more experienced than their younger colleagues cannot profit more from practice on the job.

9.4 EXPERIMENT 2

As could be shown in Experiment 1, there is a correlation of age on detection performance in the X-Ray ORT, a speeded visual search test. The aim of Experiment 2 was to replicate the age effect on X-ray screening tasks with another sample of screeners and examine whether similar effects can be found when using an on-the-job measure of X-ray detection performance. Furthermore, it can be investigated if age effects differ depending on the task demands, i.e., a frequent decision about the threat potential of an item versus the rare detection of a threat event are required. To this end, TIP data was used. TIP is a technology of current X-ray equipment that allows the projection of fictional X-ray images of threat objects into X-ray images of real passenger bags (usually ever 30-100 bags). TIP data thus provide a valid measure of on-the-job performance in X-ray screening. Again, a partial correlation between detection performance d' and years since employment should investigate whether experience tends to mitigate the effect of aging.

9.4.1 Method

9.4.1.1 Participants

In Experiment 2, 155 CBS screeners between 20 and 61 years (M = 37.92, SD = 10.34) took the X-Ray ORT and detection performance on the job was measured using TIP data over one to two years. Again, screeners were working a minimum of 12 hrs/week for at least 1.61 years (M = 5.98, SD = 3.60) and are comparable in job-related knowledge. Again, an outlier analysis was calculated and all values higher or lower than two standard deviations from the mean were excluded. Thus, five screeners were excluded for the X-Ray ORT and 11 for the TIP data analysis.

9.4.1.2 Materials and Procedure

All aviation security screeners took the X-Ray ORT in a well lit computer classroom. The X-ray screening test was run on Priminfo computers using the PCQuest software. All X-ray images were presented on full screen size at a resolution of 1024 x 768 pixels on a 17-inch TFT monitor. Brightness and contrast settings on each screen were the same, 100 and 97, respectively.

In addition, TIP data were evaluated. For CBS screeners, the TIP system displays fictional threat items into X-ray images of real passenger bags in random order (1-3% of all bags). After each TIP image screeners receive a feedback message that a fictional threat item was present so that no negative impact on screening operations occurs. A standard library based on FAA (Federal Aviation Administration) which is available on current TIP systems was used and TIP data were aggregated over a period of two years[22].

9.4.2 Results and Discussion

Overall, the results from Experiment 1 could be replicated. A partial correlation between detection performance in the X-Ray ORT and age of screeners controlling for years since employment was significant $pr = -.27$, $p < .01$. Thus, on average increased age of aviation security screeners was associated with decreased detection performance in the interpretation of X-ray images. Furthermore, Experiment 2 examined whether the long term correlation between age and detection performance can also be shown when measuring performance on the job using TIP. Again, the partial correlation between d' in TIP and age of screeners taking years of employment as control variable into account, showed a significant negative effect ($pr = -.34$, $p < .01$) and indicates that older screeners perform also worse in everyday working life when practice on the job is given. Figure 7.4 shows the correlations between detection performance d' and age of screeners in the X-Ray ORT and the TIP. The correlation of $r = -.18$, $p < .05$ between d' in the X-Ray ORT and years in age already indicates that in this sample experience in the X-Ray ORT could have a positive influence on detection performance.

[22] For CBS TIP data the number of harmless bags have to be estimated using TIP to bag ratio.

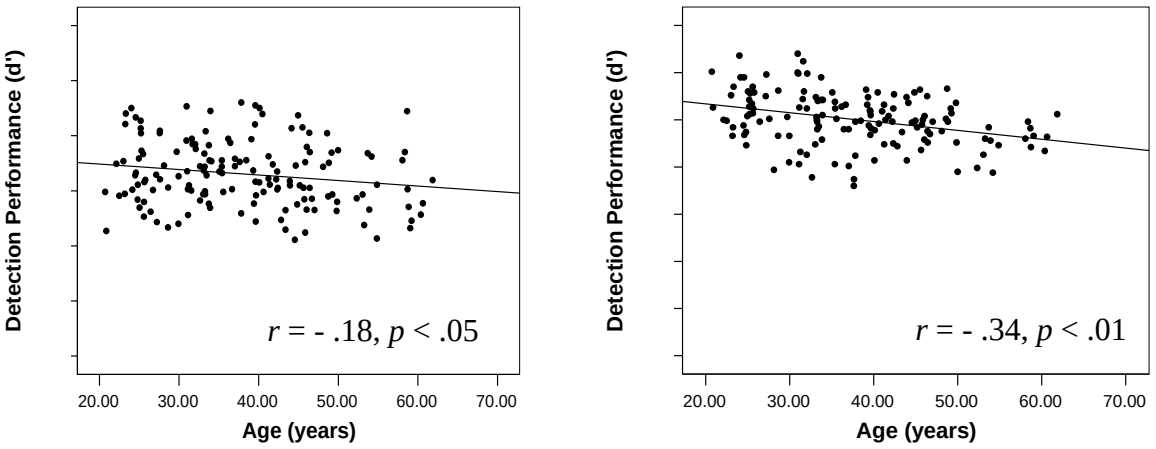

Figure 7.4. Correlation between detection performance (d') in the X-Ray ORT (left) and TIP (right).

Partial correlation between d' and years since employment controlling for age was in fact significant in the X-Ray ORT (*pr* = .35, *p* < .01), but not for TIP data (*pr* = .02, *p* = .78). Especially on the job no beneficial effect of experience on detection performance could be found. Contrary to the previous result, screeners who are employed for a long time seem to profit more from experience in the X-Ray and with sufficient amount of practice might be able to achieve similar performance levels as young adults. This effect could be due to the different selection procedure of new employees at the two airports.

Overall, the results from Experiment 1 were replicated. Furthermore, analysis of on-the-job data (TIP) showed clear evidence that the occurrence of threat items does not influence detection performance and has therefore neither a beneficial influence on performance of older or younger workers.

9.5 GENERAL DISCUSSION

In this study, we investigated the influence of age on X-ray detection performance of airport security screeners using two different X-ray screening tasks. Additional analyses showed if the age effect could be reduced by means of working experience. Overall, the result revealed that increased age was in fact associated with decreased detection performance in X-ray screening. There were clear long term correlations between d' in the X-Ray ORT and age. This effect of aging could also be found with TIP data, a measure of detection performance on the job. Further analyses on correlations between detection performance in X-ray screening and years of

employment suggest that practice on the job does generally not help to increase detection performance in both test conditions.

Considering detection performance in the X-Ray ORT, a speeded visual search task, results from both European airports show clear long term correlations between recognizing guns and knives in X-ray images of passenger bags and the age of screeners. Furthermore, this age effect could also be shown with TIP data, a more realistic task in which all kind of threat items are projected into X-ray images of real passenger bags in random order. Taken together, our results suggest stable negative age effects in a selective sample of airport security screeners that were not overcome by the amount of job experience and extensive job-specific use of speeded visual search and detection abilities.

In general correlations between detection performance and years of employment support the assumption that practice on the job does not help to increase performance in X-ray screening. However, results regarding the X-Ray ORT were not consistent. In Experiment 1 no effect could be found whereas Experiment 2 revealed a rather large correlation between the detection of guns and knives in the X-Ray ORT and years since employment. This positive correlation could be due to the different selection procedures at airports. Since years, at Airport 1 job applicants were selected with some basic X-ray images among other tests. As the X-Ray ORT measures mainly fluid abilities needed in X-ray screening it could be that due to this selection criterion, screeners at Airport 1 were better in the interpretation of X-ray images when they got employed compared to screeners of Airport 2 and thus show no effect of working experience. Furthermore, this positive effect of working experience on X-ray screening could only be shown for the X-Ray ORT. No influence of practice on TIP performance could be found.

Moreover, correlations shown in Figure 7.3 and Figure 7.4 reveal large interindividual differences. In fact, there are screeners over 60 years who perform on a remarkably higher level than some screeners that are half their age. The question whether this difference between older screeners is due to specific abilities to cope with the image-based factors or rather a guided, systematic training on the job have to be investigated in further studies. As Schwaninger et al. (2005) pointed out, the importance of knowledge-based factors in X-ray screening tasks should be taken into account. In order to recognize prohibited items in passenger bags, one has to know what they look like in X-ray images. The appearance of some objects in X-ray images is quite different than in reality (for example a teaser). Furthermore, some objects

such as improvised explosive devices (IEDs) are normally not seen at checkpoints. Thus, working experience alone is probably not sufficient to attenuate the age-related differences in X-ray screening. Previous studies in airport security could show that specific individually adaptive computer-based training increases detection of threat items in passenger bags (Koller et al., 2008; Schwaninger & Hofer, 2004). To this end aviation security screeners did not received specific computer-based training, but show a rather high amount of working experience (up to 26 years). Whether a specific job-related training can reduce the age effect has to be investigated in further studies.

In general, most of our findings are consistent with earlier findings on age-related changes in cognition across middle adulthood. However, this study on age effects in airport security provides further insights into area specific applications and transferability of research studies into real world conditions. Our result about the domain specific age effect is in agreement with previous studies which could show that age influences different processes such as visual search, spatial imagination, working memory and attention. However, overall our results are not consistent with studies showing an experience-related increase in performance. Zec (1995) for example reported that experienced individuals adopt more efficient strategies and, thus, increase their performance. It could also be assumed that older screeners might compensate their deficits with their knowledge, working experience and probably working strategy. However, with one exception there were no significant partial correlations between detection performance (X-Ray ORT and TIP) and years since employment when age was controlled for. Thus, in general aviation security screeners cannot compensate their decline in job specific fluid abilities with higher working experience. This holds true for both types of tasks examined as detection performance in TIP did also not increase with more working experience. However, further studies should investigate if older worker can compensate a negative age effect through a reflected job experience such as systematic adaptive training systems.

10. TRAIN IT OR LOSE IT: THE INFLUCENCE OF AGE AND TRAINING ON DETECTION PERFORMANCE IN X-RAY SCREENING

10.1 ABSTRACT

X-ray screening of passenger bags is a highly demanding task which requires visual cognition abilities and the knowledge about the appearance of threat items in X-ray images. A previous study has shown that age of screeners has a negative effect on the detection performance in X-ray screening which could not be compensated with working experience. As X-ray screening strongly depends on training rather than working experience alone, the influence of training compared to working experience on both, visual cognition abilities and knowledge in X-ray screening was investigated. It was found that the age effect regarding image-based factors remains similar before and after two years of training. However, for knowledge-based factors an even larger age effect could be found. Learning effects for the detection of all kinds of prohibited items and the interaction between the ability to cope with image-based factors and the knowledge in X-ray screening were discussed. Further, the question whether training would compensate age related declines if older screeners have had more training according to their age, have to be investigated in future research studies[23].

10.2 INTRODUCTION

Research on cognitive aging has shown a negative relation between age and cognitive performance for different tasks. Not only specific processes such as spatial imagination and visual search are affected by aging (Dror & Kosslyn, 1994; Campos et al., 2004; Humphrey & Kramer, 1997; Kramer & Atchley, 2000; Laberge and Scialfa, 2000), but also more general structures and processes used for temporarily storing and manipulating information. For example Salthouse and Babcock (1991) as well as Cherry and Park (1993) reported a decrease of working memory with aging. Further, attention which implies withdrawal from some things in order to deal effectively with others is often impaired when people grow older (Roger & Fisk,

[23] I gratefully acknowledge the help of Mike Martin and Adrian Schwaninger in designing the study.

2001). A negative effect of age was also found in the field of aviation security for the X-ray screening task. Schwaninger, Hardmeier, Riegelnig, and Martin (in preparation) revealed a rather large age effect for the detection of threat items in X-ray images of passenger bags. These declines could be attributed to the required visual cognition processes in X-ray screening. A screener has to recognize prohibited items regardless of high bag complexity, superposition and rotation (image-based factors). Therefore, age-related visual cognition processes such as visual search, figure-ground segregation and mental rotation are needed. Whereas previous studies could have shown that such age declines on the job can be compensated quite often by working experience (Zec, 1995; Kliegl et al., 1989), no decrease for age related declines with working experience could be found for the X-ray screening task (Schwaninger et al., in preparation). As the detection performance in X-ray screening is influenced significantly by training, working experience alone is probably not sufficient to reduce the age-related differences. According to Schwaninger and Hofer (2004), Schwaninger (2005), Koller et al. (2008) detection performance in X-ray screening can be increased significantly with an individual adaptive training system as the appearance of prohibited items in X-ray images often differs remarkably from the appearance in reality. Items like electric shock devices, self-defense gas-sprays or plastic pistols look quite different in reality and without specific training they are hardly ever recognizable in an X-ray image. Again other items like improvised explosive devices (IEDs) are normally not seen at checkpoints and have therefore to be memorized with a training system. As well Hardmeier et al. (2006b) found a significantly increase in detection performance for experienced aviation security screeners after two years of individually adaptive training. Thus, results imply that experience is probably not enough to store the visual appearance of all kinds of prohibited items in the visual memory. Especially considering that some threat items like IEDs are normally not seen at checkpoints. A training system enables considerably more exposure to threat items linked with direct feedback compared to the one on the job.

In this study we investigated the influence of age and training on detection performance for both, image-based and knowledge-based factors. Image-based factors were defined by Schwaninger et al. (2005) and refer to the ability to cope with bag complexity, superposition and rotation of threat items in X-ray images. In contrast to image-based factors, knowledge-based factors show whether a screener knows which items are prohibited and what they look like in X-ray images. To

measure these factors relatively independent of each other two X-ray screening tests were used. The X-Ray Object Recognition Test (X-Ray ORT) measures image-based factors relatively independent of knowledge as only guns and knives are used. The Prohibited Items Test (PIT) includes all kinds of prohibited items in an easy view whereas bag complexity and superposition were kept relatively constant and thus measures mainly knowledge-based factors. First, the influence of age on detection performance in both tests was investigated before and after two years of training. Based on the previous study, we expect a negative effect of age on detection performance for both tests. Whether the detection performance increase for image-based and knowledge-based factors differs for older and younger screeners remains to be shown. As image-based factors are assumed to be relatively stable abilities, the increase after training (which is expected to be rather small) should be comparable for younger and older screeners. However, in the PIT different performance increases are likely. Based on these findings, further analyses should investigate whether the learning effect depends on the baseline performance and whether an interaction between ability of screeners and performance increase can be observed. Again, the influence of age on these learning effects is discussed.

10.3 METHOD

10.3.1 Participants

The data for this study was collected as part of the recurrent verification of screening competency for aviation security screeners. Data from 334 screeners (101 male and 233 female) who took both tests in 2004 and 2006 was used. Ages ranged from 23 to 62 years (M = 46.71, SD = 8.37) and mean working experience was 7.54 years (SD = 5.13, range: 1 to 23 years) when taking the first test run in 2004. Between the two measurements screeners had on average two times 20 minutes individually adaptive CBT with X-Ray Tutor (Number of logins: M = 207.88, SD = 112.12). For the analyses values higher or lower than 3 standard deviations from the mean were excluded (for the X-Ray ORT five and for the PIT four screeners).

10.3.2 Materials and Procedure

X-Ray Object Recognition Test (X-Ray ORT)

Image-based factors in X-ray screening were measured using the X-Ray ORT which is a reliable and valid X-ray screening test to identify the ability to cope with bag

complexity, superposition and viewpoint of threat items relatively independent of training. Therefore, only guns and knives are used in this test and all X-ray images are displayed in grayscale only. All eight guns and eight knives in the test are shown from two different viewpoints (easy and difficult). Each view is then combined with two bags of low and two bags of high complexity level once with low and once with high superposition. Thus, in total each gun and each knife is shown eight times in different conditions. The test includes a total of 256 trials: 2 weapons (guns and knives) * 8 (exemplars) * 2 (views) * 2 (superpositions) * 2 (bag complexities) * 2 (harmless vs. threat image).

The X-Ray ORT is a computer-based test in which each X-ray image is displayed for 4 seconds on the screen. Then participants have to give the answer whether the bag was OK (contained no gun or knife) or NOT OK (contained a gun or knife) clicking on the respective button on the screen. Furthermore, they have to indicate how sure they are in their decision clicking on a 90 point rating scale. The test starts with a self-explanatory instruction including eight exercise trials to familiarize the participants with the test taking procedure. In this phase participants receive a feedback whether their answer was correct or not showing the solution. In the test itself no feedback was given to participants anymore. As the X-Ray ORT should measure the visual abilities to cope with the three image-based factors and not the knowledge of test participants, all eight guns and eight knives either in the frontal or rotated view were shown for 10 seconds to them before the test started.

The test is subdivided into four blocks and participants have the possibility to take a short break after each block. The test takes about 45 minutes to complete. For more details about the X-Ray ORT as well as its reliability and validity measures see Schwaninger et al. (2005) respectively Hardmeier et al. (2006a).

Prohibited Items Test (PIT)

Knowledge-based factors in X-ray screening were measured with the PIT. The PIT includes all kinds of prohibited items in X-ray images. All threat items in the PIT can be classified into seven categories according to ICAO, ECAC and EU prohibited items lists. Thus, a total of 19 guns, 27 sharp objects, 14 hunt and blunt instruments, 5 highly inflammable substances, 17 improvised explosive devices (IEDs), 3 chemicals and 13 other prohibited items (such as self-defense gas spray) were used. All threat items are shown in an easy view. For all trials bag complexity and superposition were kept relatively constant. The PIT includes a total of 160 trials, half of them are

harmless bags, the other half include at minimum one prohibited item. 68 of all threat images include one prohibited item only and the remaining 12 trials include half two and half three prohibited items at once[24].

Again, participants had to decide whether the bag was OK (included no prohibited item) or NOT OK (included one or more prohibited items) clicking on the respective button on the screen. Furthermore, they had to indicate to which of the threat category the prohibited item(s) belong(s) to and how sure they were in their decision clicking on a 50 point rating scale[25]. Images were displayed for maximum 10 seconds on the screen. For the answer no time limit was given. By clicking the space bar the next image could be shown. Like the ORT, the PIT is computer-based and includes an introduction and some exercise trials that participants get used to the test taking procedure. Contrary to the test condition, participants were given a feedback about the correct answer for the exercise trials. The PIT takes about 45 minutes to complete and includes four blocks. After each block participants had again the possibility to take a short break. For reliability and validity measures of this test please see Hardmeier et al. (2006a).

X-Ray Tutor (XRT)

The X-Ray Tutor (XRT) is an individually adaptive training system for aviation security screeners. Screeners see X-ray images for 15 seconds on the screen and have to answer whether the bag is OK (included no prohibited item) or NOT OK (included a prohibited item). Then they receive an immediate feedback whether their answer was correct or not. Further, screeners can review any X-ray image to learn where the threat item was located. A threat information window is displayed showing the threat items in both the X-ray image and in reality (photograph). The training system contains 500 threat items in six basic views which are additionally mirrored (4x) and plane rotated (3x) and 6000 X-ray images of passenger bags. Threat items are combined with passenger bags based on the individually adaptive training algorithm of XRT. This algorithm provides screeners with X-ray images adapted in their difficulty to screeners' individual performance. For more information about X-Ray Tutor or its effectiveness, see Schwaninger (2005) and Koller et al. (2008).

[24] Only trials including one prohibited item were analyzed.
[25] For analysis only OK and NOT OK answers were used.

Procedure

All participants took both tests as part of the recurrent verification of screening competency which included a total of four tests. All screeners completed first the X-Ray ORT followed by the PIT, the bomb detection test (BDT) and a computer-based questionnaire (CBQ). The last two tests are not part of this study. Tests were taken in a well lit computer-classroom including 10 HP computers. The test was run on a 17-inch monitor set at a resolution of 1024 x 768 pixels. Participants used a standard key-board and a "Microsoft Optical Wheel Mouse" to give the answers.

10.4 RESULTS

Results in both tests were calculated using the detection performance measure d'. D' is a psychophysical measure which takes into account the hit-rate and the false alarm rate. A good screener should reach a high hit rate without rejecting too many harmless bags. For more information about d' see Green and Swets (1966), MacMillan and Creelman (1991).

The result section is organized as follows. First, effects of age on image-based and knowledge-based factors before and after two years of training are reported. Then training effects for the PIT were investigated and discussed in reference to young and elderly screeners.

10.4.1 Effect of age on image-based and knowledge-based factors

Figure 8.1 provides the correlations between detection performance and age of screeners. Partial correlation between d' and age of screeners taking working experience into account revealed a significant effect for the X-Ray ORT $pr = -.27$, $p < .01$ and a significant effect for the PIT $pr = -.22$, $p < .01$ at the baseline measurement (2004). After two years of training (2006) partial correlations controlling for working experience and training hours revealed again a significant negative partial correlation for the X-Ray ORT $pr = -.34$, $p < .01$ and a even larger one for the PIT $pr = -.40$, $p < .01$. Thus, the results show a decrease in the detection of prohibited items with age for both tests before and after two years of training with XRT. Whereas the age effect seems to remain relatively stable across both measurement conditions in the X-Ray ORT, it is clearly larger after training regarding the PIT (see Figure 8.1c and 8.1d).

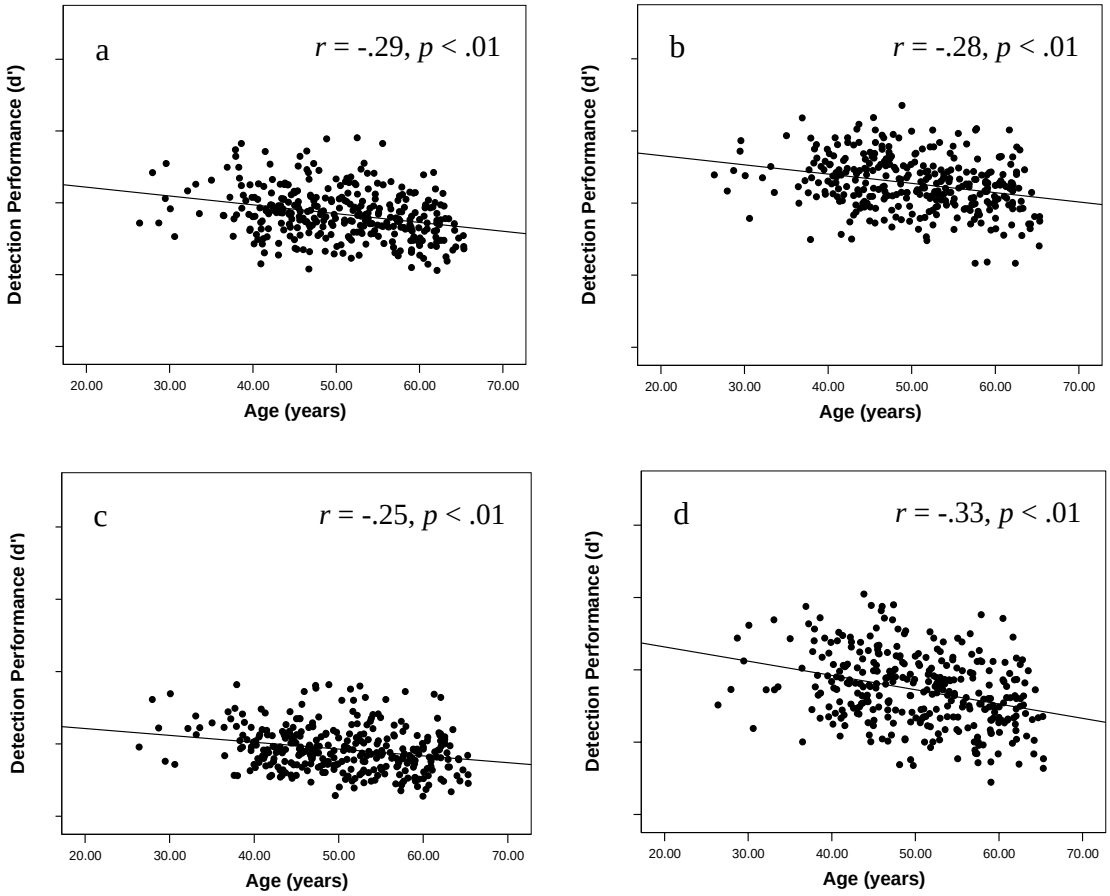

Figure 8.1. Correlation between detection performance (d') and age for (a) the X-Ray ORT 2004, (b) the X-Ray ORT 2006, (c) the PIT 2004 and (d) the PIT 2006.

To test whether the performance increase before and after two years of training is similar for younger and older screeners, an analysis of covariance (ANCOVA) with measurement (2004, 2006) as within-participant factor, age (Q1, Q4) as between-participant factor and training and years since employment as covariate was conducted (see also Figure 8.2). For the between-participant factor age the first quartile (M = 39.94, SD = 4.90) and last quartile (M = 57.24, SD = 2.11) was calculated and used for analysis[26]. Taking training and working experience into account, results showed a significant main effect of measurement $F(1, 157) = 19.83$, $p < .01$, $\eta^2 = .11$, a significant main effect of age $F(1, 157) = 44.40$, $p < .01$, $\eta^2 = .22$, and no significant interaction between measurement and age $F(1, 157) = 0.00$, $p = .97$ for the X-Ray ORT. Further, both interactions between measurement and the covariates were not significant (p > .07). However, the ANCOVA for the PIT

[26] Results showed similar effects calculating a median split instead of quartiles.

revealed a significant main effect of measurement $F(1, 157) = 72.34$, $p < .01$, $\eta^2 = .32$, a significant effect of age $F(1, 157) = 42.52$, $p < .01$, $\eta^2 = .21$ *and* a significant interaction between measurement and age $F(1, 157) = 13.52$, $p < .01$, $\eta^2 = .08$. Interactions between measurement and training as well as working experience were also significant $F(1, 157) = 17.34$, $p < .01$, $\eta^2 = .10$ and $F(1, 157) = 6.28$, $p < .05$, $\eta^2 = .04$ respectively. Thus, an increased detection performance could be found for both tests due to the training system. The ability to cope with image-based factors

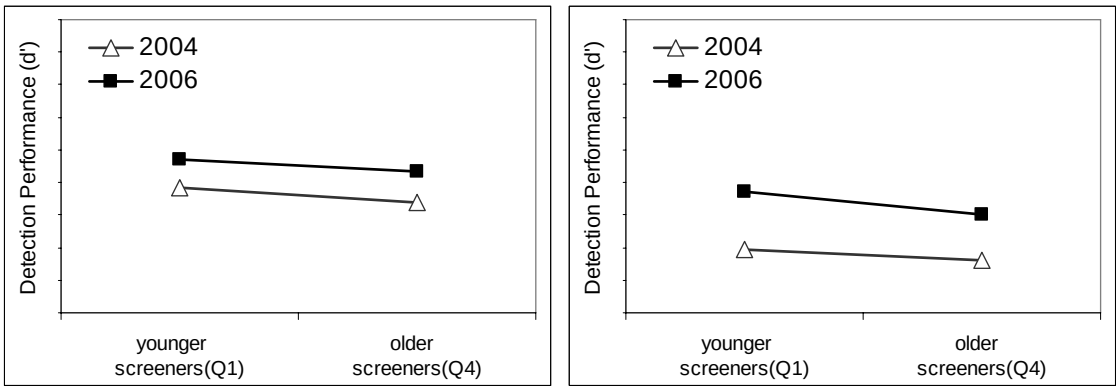

Figure 8.2. Detection performance in the X-Ray ORT (left) and the PIT (right) for the youngest (Q1) and the oldest (Q4) screeners.

and the knowledge about prohibited items was worse for older screeners compared to younger ones before and after two years of training. Whereas detection performance increase with training was similar for younger and older screeners regarding to image-based factors, it was different in the PIT. Performance increase of older screeners was much smaller than the one of younger ones although on average they trained significantly more ($t(159) = -3.,92$ $p < .01$). However it has to be noted that the correlation between age and training was $r = .27$, $p < .01$ only.

Generally, the large individual differences in detection performance in the PIT after two years of training (see also Figure 8.1d) imply that differences between screeners increase in this first training phase. Thus, it can be assumed that some screeners learn more and probably faster than their colleagues.

10.4.2 Learning effect in the PIT

Interestingly, detection performance of both, younger and older screeners varied enormously after two years of training in the PIT (Figure 8.1d). Controlling for working experience and training we found a rather large relationship between

detection performance at the baseline measurement and the relative increase[27] in detection performance r = - .64 (p < .01). Similar partial correlations were found for younger (r = - .69, p < .01) and older screeners (r = - .65, p < .01). Consistent with previous findings and learning theories the training effect is larger for screeners who perform poorly at the beginning.

Furthermore, we investigated whether screeners who perform on a high level in the X-Ray ORT and thus have the ability to cope with image-based factors perform as well better in the PIT and show a larger training effect compared to people with poor detection performance in the X-Ray ORT. Low and high performer in the X-Ray ORT were defined with a median split[28]. An ANCOVA with the within-participant factor measurement (2004, 2006) and the between-participant factor ability (low vs. high performer) controlling for age, working experience and training revealed a significant main effect of measurement F(1, 315) = 66.77, p < .01, η^2 = .18 and ability F(1, 315) = 77.61, p < .01, η^2 = .20, but no significant interaction between measurement and ability F(1, 315) = 1.25, p = .27. All interactions between measurement and the covariates were also significant p < .01, η^2 > .03 Thus, screeners with high visual cognition abilities perform as well better in the PIT before and after two years of training. However, the increase in detection performance was similar for both groups. The ability to cope with image-based factors has no positive effect on the learning capacity (see also Figure 8.3). As can be seen in Figure 8.3 similar results were found for

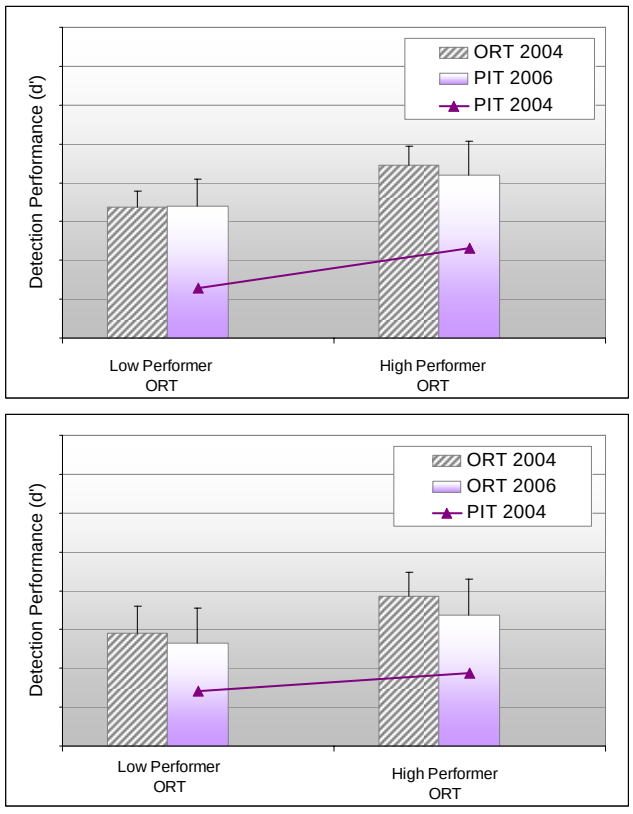

Figure 8.3. Detection performance in the PIT before and after two years of training for low and high performers in the X-Ray ORT. On the top for younger screeners and below for older screeners.

[27] % Relative Increase in d' = (m2 - m1)/m1 (m1 means first measurement and m2 means second measurement).

[28] Results using quartiles instead of the median showed similar effects. We decided to use a median split in order to have an enough large sample size available when calculating the ANCOVA for younger and older screeners only.

younger (Q1) and older screeners (Q4) (see Table 5.1). Although the main effect of measurement was not significant for both groups, Bonferroni-corrected pairwise comparisons between the two measurements showed significant effects ($p < .01$) confirming training effectiveness. Again, there was a significant effect of ability, but no significant interaction between measurement and ability implying that despite different abilities the learning effect within these two years was similar.

Table 5.1

Results of the ANCOVA for both, younger and older screeners

	Factor	*df*	*F*	*p*	η^2
Younger screeners (Q1) N = 80	Measurement	75	1.53	.22	.02
	Ability	75	27.68	< .01	.27
	Measurement * Ability	75	0.68	.41	.01
	Measurement * Age	75	0.88	.35	.01
	Measurement * Experience	75	7.35	< .01	.09
	Measurement * Training	75	10.94	< .01	.13
Older screeners (Q4) N = 80	Measurement	75	2.72	.10	.04
	Ability	75	15.27	< .01	.17
	Measurement * Ability	75	0.71	.40	.01
	Measurement * Age	75	1.39	.24	.02
	Measurement * Experience	75	1.17	.28	.02
	Measurement * Training	75	4.22	< .05	.05

10.5 DISCUSSION

The main objective of this study was to investigate whether the age effect in X-ray screening is similar for visual cognition processes and the knowledge in X-ray screening respectively. Further, we examined whether this age effect remains similar after two years of training. We could show that age-related declines can be seen for both tasks before and after two years of training with an individually adaptive training system. Thus, neither experience nor training can compensate age related declines in this study. These results are not consistent with previous studies which showed that older workers can compensate age-related declines quite often with working experience (Zec, 1995; Kliegl et al., 1989). These contradictory findings can be explained as follows. The influence of working experience for the X-ray screening

task is relatively small as learning effects on the job are limited. On the one hand a screener sees only some prohibited items in everyday work. For example IEDs are normally not seen at checkpoints. On the other hand, most prohibited items that are brought along by passenger are too seldom seen to form a visual memory representation of these objects. Thus, with working experience alone a screener is not able to familiarize with all possible prohibited items needed to represent all different types of threat items in the visual memory (Hardmeier et al., 2006b; Koller et al., 2008). As the training system in this study was implemented in 2004 and therefore older screeners cannot profit more from the training due to their longer employment smaller training effects for older screeners could have been expected. Many studies report impaired visual cognition processes with age which can be assumed to be relevant for the learning process (Salthouse & Babcock, 1991; Deaton & Parasuraman, 1993). However, future studies should investigate whether training can help to reduce the age effect in X-ray screening if screeners who are employed since years have accordingly more training. Moreover, results from longitudinal studies would be interesting in order to find out how performance changes over the years for screeners employed in younger days. Interestingly, results showed more training hours for older screeners which are generally employed for a longer time implying that older screeners try to compensate their decline with more training. Contradictory results were found related to memory awareness. However, many studies reported that older people rather overestimate their performance than vice versa (Bruce, Coyne, & Botwinick, 1982; Murphy, Sanders, Gabriesheski, & Schmitt, 1981). Based on the results of this study it can be assumed that this overestimation of one's own capabilities is probably reduced in areas requiring high expertise.

Results also revealed that the learning effect in the PIT is smaller for screeners who perform already on a relatively high level compared to screeners with poor detection performance at the beginning. This can be expected as learning effects are generally larger at the beginning and decrease with increasing expertise. This effect can be seen for younger and older screeners.

Further we investigated whether an interaction between the ability of screeners and the learning effect could be found. Results showed a similar performance increase in the PIT for people with low and high abilities to cope with image-based factors. In this initial training stage it does not matter whether people are able to cope with image-based factors or not. However, it should be noted that this fact does not reduce the importance of visual cognition abilities. First, screeners with good abilities

perform on a significantly higher level. Second, prohibited items in the PIT were shown in an easy view and with medium superposition and bag complexity. However, in real life these factors always play along and should be considered. Third, it could be argued that the same increase of able screeners who performed already on a higher level is relatively more valuable. Interestingly, the learning effect of older screeners differs not from the one of younger screeners.

In summary, this study clearly showed the relatively large effect of age on detection performance in X-ray screening as well after training. However, it remains to be shown whether older screeners could compensate age-related declines if their training would match their working experience or in other words if their training experience would have started when they got employed.

11. REFERENCES

Amthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (2001). *Intelligenz-Struktur-Test 2000 R* (*I-S-T 2000 R), Manual.* Göttingen: Hogrefe.

Arbuckle, J. L. (2005). *Amos 6.0 users guide*. Chicago, IL: SPSS.

Baddeley, A. (1986). *Working memory*. New York: Oxford University Press.

Baldwin, T., & Ford, J. K. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology, 41*, 63-105.

Bentler, P. M. (1992). On the fit of models to covariances and methodology to the Bulletin. *Psychological Bulletin, 112*, 400-404.

Berg, C., Hertzog, C., & Hunt, E. (1982). Age differences in the speed of mental rotation. *Developmental Psychology, 18(1),* 95-107.

Biederman, I. (1987). Recognition-By-Components: a theory of human image understanding. *Psychological Review, 94*, 115-147.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Borkenau, P., & Ostendorf, F. (1993). *NEO-Fünf-Faktoren-Inventar (NEO-FFI) nach Costa und McCrae, Manual.* Goettingen: Hogrefe.

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences of the United States of America, 89*, 60-64.

Bülthoff, H. H., Edelman, S. Y., & Tarr, M. J. (1995). How are three-dimensional objects represented in the brain? *Cerebral Cortex, 3*, 247–260.

Byrne, B. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. New Jersey: Erlbaum.

Byrne, B., Shavelson, R. J., & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105(3)*, 456-466.

Cabeza, R. (2001). Functional neuroimaging of cognitive aging. In R. Cabeza & A. Kingston (Eds.), *Handbook of functional neuroimaging of cognition*. Cambridge: MIT Press, MA.

Campos, A., Pérez-Fabello, M. J., & Gómez-Juncal, R. (2004). Gender and age differences in measured and self-perceived imaging capacity. *Personality and Individual Differences*, *37,* 1383-1389.

Cherry, K. E., & Park, D. C. (1993). Individual difference and contextual variables influence spatial memory in younger and older adults. *Psychology and Aging, 8,* 517-526.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Erlbaum, Hillsdale.

Deaton, J. El., & Parasuraman, R. (1993). Sensory and cognitive vigilance: Effects of age on performance and subjective workload. *Human Performance*, *6,* 71-97.

Dror, I. E., & Kosslyn, S. M. (1994). Mental imagery and aging. *Psychology and Aging*, *9(1)*, 90-102.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation, 3*, 194-200.

Fozard, J. L., & Gordon-Salant, S. (2001). Changes in vision and hearing with aging. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging.* San Diego: Academic Press.

Ghylin, K. M., Drury, C. G., & Schwaninger, A. (2006). Two-component model of security inspection: application and findings. *16th World Congress of Ergonomics, IEA 2006*, *Maastricht, The Netherlands, July, 10-14, 2006*.

Graf, M., Schwaninger, A., Wallraven, C., & Bülthoff, H.H. (2002). Psychophysical results from experiments on recognition & categorisation. *Information Society Technologies (IST) programme,* Cognitive Vision Systems - CogVis (IST-2000-29375).

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin, 75*, 424-429.

Hampson, E. (1995). Spatial cognition in humans: Possible modulation by androgens and estrogens. *Journal of Psychiatry and Neuroscience, 20(5),* 397-404.

Hardmeier, D., Hofer, F., & Schwaninger, A. (2005). The X-Ray Object Recognition Test (X-Ray ORT) – A reliable and valid instrument for measuring visual abilities needed in X-ray screening. *IEEE ICCST Proceedings*, *39*, 189–192.

Hardmeier, D., Hofer, F., & Schwaninger, A. (2006a). Increased detection performance in airport security screening using the X-Ray ORT as pre-employment assessment tool. *Proceedings of the 2nd International Conference on Research in Air Transportation, Belgrade, Serbia and Montenegro, June 24-28, 2006,* 393-397.

Hardmeier, D., Hofer, F., & Schwaninger, A. (2006b). The role of recurrent CBT for increasing aviation security screeners' visual knowledge and abilities needed in X-ray screening. *Proceedings of the 4th International Aviation Security Technology Symposium, Washington, D.C., USA, November 27 – December 1, 2006,* 338-342.

Hayward, W. G. (2003). After the viewpoint debate: where next in object recognition? *Trends in Cognitive Sciences, 7(10),* 425–427.

Hofer, F., & Schwaninger, A. (2004). Reliable and valid measures of threat detection performance in X-ray screening. *IEEE ICCST Proceedings, 38*, 303-308.

Hofer F., & Schwaninger, A. (2005a). Using threat image projection data for assessing individual screener performance. *WIT Transactions on the Built Environment, 82,* 417-426.

Hofer F., & Schwaninger, A. (2005b). Using threat image projection data for assessing individual screener performance. In C.A.Brebbia, T. Bucciarelli, F. Garzia, & M.Guarascio, *WIT Transactions on the Built Environment (82), Safety and Security Engineering* (pp. 417-426). Wessex: WIT Press.

Horn, W. (1983). *Leistungsprüfungssystem (L-P-S). Handanweisung für die Durchführung, Auswertung und Interpretation.* 2., erweiterte und verbesserte Auflage. Göttingen: Hogrefe.

Howell, W. C., & Cooke, N. J. (1989). Training the human information processor: A look at cognitive models. In I. L. Goldstein (Ed.), *Training and development in work organizations: Frontiers of industrial and organizational psychology* (pp. 121-182). San Francisco: Jossey-Bass.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1-55.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review, 99(3)*, 480-517.

Humphrey, D. G., & Kramer, A. F. (1997). Age differences in visual search for feature, conjunction, and triple-conjunction targets. *Psychology and Aging, 12,* 704-717.

Ishihara, S. (2005). *Ishihara's test chart for colour deficiency.* Tokyo: Kanehara trading Inc.

Janowsky, J. S., Oviatt, S. K., & Orwoll, E. S. (1994). Testosterone influences spatial cognition in older men. *Behavioral Neuroscience, 108(2),* 325-332.

Jonassen, D. H., Hannum, W. H., & Tessmer, M. (1989). *Handbook of task analysis procedures*. New York: Praeger.

Kirwan, B., & Ainsworth, L. K. (1992). *A guide to task analysis*. London: Taylor & Francis.

Kliegl, R., Smith, J., & Baltes, P. B. (1989). Testing-the-limits and the study of adult age differences in cognitive plasticity of a mnemonic skill. *Developmental Psychology, 25,* 247-256.

Kline, P. (2000). *The handbook of psychological testing*. London: Routledge.

Koller, S., & Schwaninger, A. (2006). Assessing X-ray image interpretation competency of airport security screeners. *Proceedings of the 2nd International Conference on Research in Air Transportation, Belgrade, Serbia and Montenegro, June 24-28, 2006*, 399-402.

Koller, S. M., Hardmeier, D., Michel, S., & Schwaninger, A. (2008). Investigating training, transfer and viewpoint effects resulting from recurrent CBT of X-ray image interpretation. *Journal of Transportation Security, 1, 81-106.*

Kosslyn, S. M. (1994). *Image and brain: The resolution of the imagery debate*. Cambridge, Massachusetts: MIT Press.

Kramer, A. F., & Atchley, P. (2000). Age effects in the marking of old objects in visual search. *Psychology and Aging, 15,* 286-296.

Laberge, J. C., & Scialfa, C. T. (2000). Predictors of web navigation performance in a life span sample of adults. *Human Factors*, *47,* 289-302.

Lawson, R. (1999). Achieving visual object constancy across plane rotation and depth rotation. *Acta Psychologica, 102,* 221–245.

Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience, 19,* 577–621.

Lowe, D. G. (1985). Perceptual organization and visual recognition. Boston: Kluwer Academic Publishing.

Lowe, D. G. (1987). Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence, 3,* 355-395.

MacMillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. Cambridge: University Press.

Maguire, E. A., Gadian, D. G., & Johnsrude, I. S. (2000). *Proceedings of the National Academy of Sciences of the United States of America, 97,* 4398-4403.

Martin, M., & Zimprich, D. (2005). Cognitive development in midlife. In S. L. Willis & M. Martin (Eds.), *Middle adulthood: A lifespan perspective* (pp. 179-206). Thousand Oaks, CA: Sage.

McCarley, J. S., Kramer, A. F., Wickens, C. D., Vidoni, E. D., & Boot, W. R. (2004). Visual skills in airport screening. *Psychological Science, 15(5)*, 302-306.

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B, Containing papers of a Biological character. Royal Society (Great Britain), 200,* 269–294.

Marsh, H. W., Hau, K. T. & Wen, Z. (2004). In search of golden rules: Comment on hypothesis testing approaches to setting cutoff values for fit indexes and dangers in overgeneralising Hu & Bentler's (1999) findings. *Structural Equation Modelling, 11*, 320-341.

Marxer, P. (2004). *Visual tests for X-ray screening*. Unpublished master's thesis, University of Zurich, Zurich, Switzerland.

Moosbrugger, H., & Oehlschlägel, J. (1996). *Frankfurter Aufmerksamkeits-Inventar: FAIR, Testmanual.* Bern: Huber.

Palmer, S. E. (1999). *Vision science – photons to phenomenology.* Cambridge, Massachusetts: the MIT Press.

Palmer, S. E., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In I. Long & A. Baddeley (Eds.), *Attention and performance* IX. Hillsdale, N.J: Erlbaum.

Palmeri, T. J., & Gauthier, I. (2004). Visual object understanding. *Nature Review Neuroscience, 5,* 291–303.

Pastore, R. E., Crawley, E. J., Berens, M. S., & Skelly, M. A. (2003). Nonparametric" A' and other modern misconceptions about signal detection theory. *Psychonomic Bulletin & Review, 10*, 556-569.

Peissi, J., & Tarr, M.J. (2007). Visual object recognition: Do we know more now than we did 20 years ago? *Annual Review of Psychology, 58,* 75–96.

Poggio, T., & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature, 343(6255)*, 263-266.

Pollack, I., & Norman, D. A. (1964). A non-parametric analysis of recognition experiments. *Psychonomic Science, 1*, 125-126.

Raven, J.C., Court, J., & Raven, J. Jr. (1980). *RAVEN-Matrizen-Test. Manual, deutsche Bearbeitung von Heinrich Kratzmeier unter Mitarbeit von Ralf Horn, Manual.* Weinheim: Beltz Test Gesellschaft.

Riegelnig, J., & Schwaninger, A. (2006). The influence of age and gender on detection performance and the criterion in X-ray screening. *Proceedings of the 2nd International Conference on Research in Air Transportation, Belgrade, Serbia and Montenegro, June 24-28, 2006,* 403-407.

Roger, W. A., & Fisk, A. D. (2001). Understanding the role of attention in cognitive aging research. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging.* San Diego: Academic Press.

Saks, A. M., & Belcourt, M. (2006). An investigation of training activities and transfer of training in organizations. *Human Resource Management, 45(4),* 629-648.

Salthouse, T. A. (1996). The processing speed theory of cognitive aging. *Psychological Review, 103,* 403-428.

Salthouse, T. A., & Babcock, R. L. (1991). Decomposing adult age differences in working memory. *Developmental Psychology, 27,* 763–776.

Schaarschmidt, U., & Fischer, A. (1996). *AVEM - Arbeitsbezogenes Verhaltens- und Erlebensmuster, Handanweisung*. Frankfurt: Swets & Zeitlinger.

Schaie, K. W. (2005). *Developmental influences on adult intelligence: The Seattle longitudinal study.* Oxford: University Press.

Schwaninger, A. (2003a). Detection systems: Screener evaluation and selection, *AIRPORT*, *2*, 14-15.

Schwaninger, A. (2003b). Training of airport security screeners. *AIRPORT,* 05/2003, 11-13.

Schwaninger, A. (2004). Computer based training: a powerful tool to the enhancement of human factors. *Aviation Security International*, FEB/2004, 31-36.

Schwaninger, A. (2005a). Object recognition and signal detection. In B. Kersten (Ed.), *Praxisfelder der Wahrnehmungspsychologie* (pp. 108-132). Bern: Huber.

Schwaninger, A. (2005b). Increasing efficiency in airport security screening. *WIT Transactions on the Built Environment, 82,* 405-416.

Schwaninger, A. (2005c). X-ray imagery: enhancing the value of the pixels. *Aviation Security International,* Oct 2005, 16-21.

Schwaninger, A., Hardmeier, D., & Hofer, F. (2005). Aviation security screeners visual abilities & visual knowledge measurement. *IEEE Aerospace and Electronic Systems, 20(6)*, 29-35.

Schwaninger, A., Hardmeier, D., Riegelnig, J., & Martin, M. (2008). *Use it and still lose it: The influence of age and job experience on detection performance in X-ray screening*. Manuscript in preparation.

Schwaninger, A., & Hofer, F. (2004). Evaluation of CBT for increasing threat detection performance in X-ray screening. In K. Morgan and M. J. Spector (Eds.), *The Internet Society 2004, Advances in Learning, Commerce and Security*. Wessex : WIT Press.

Schwaninger, A. & Hofer, F. (2004). Evaluation of CBT for increasing threat detection performance in X-ray screening. In K. Morgan & M. J. Spector, *The internet society 2004, advances in learning, commerce and security* (pp. 147-156). Wessex: WIT Press.

Schwaninger, A., Hofer, F., & Wetter, O. E. (2007). Adaptive computer-based training increases on the job performance of X-ray screeners. *IEEE ICCST Proceedings, 41*, 117-124.

Schwaninger, A., Michel, S., & Bolfing, A. (2007). A statistical approach for image difficulty estimation in X-ray screening using image measurements. *Proceedings of the 4th Symposium on Applied Perception in Graphics and Visualization, ACM Press, New York, USA,* 123-130.

Scialfa, C. T., Thomas, D. M., & Joffe, K. M. (1994). Age differences in the useful field of view: An eye movement analysis. *Optometry and Vision Science, 71,* 736–742.

Seamster, T. L., Redding, R. E., & Kaempf, G. L. (1997). *Applied cognitive task analysis in aviation*. Aldershot: Avebury aviation.

Smith, J. D., Redford, J. S., Gent, L. C., & Washburn, D. A. (2005). Visual search and the collapse of categorization. *Journal of Experimental Psychology: General, 134(4),* 443-460.

Smith, J. D., Redford, J. S., Washburn, D. A., & Taglialatela, L. A. (2005). Specific-token effects in screening tasks: possible implications for aviation security. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31(6),* 1171-1185.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research, Instruments, & Computers*, *31(1)*, 137–149.

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in crossnational consumer research. *Journal of Consumer Research, 25*, 78-90.

Stryker, M. P. (1991). Temporal associations. *Nature, 354,* 108–109.

Tarr, M. J. (1995). Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin and Review, 2,* 55-82.

Tarr, M. J., & Bülthoff, H. H. (1995a). Is human object recognition better described by geon structural descriptions or by multiple views? *Journal of Experimental Psychology, Human Perception and Performance*, *21(6)*, 1494-1505.

Tarr, M. J., & Bülthoff, H. H. (1995b). Is human object recognition better described by geon structural descriptions or multiple views? Comment on Biederman and Gerhardstein (1993). *Journal of Experimental Psychology: Human Perception and Performance, 21,* 1494-1505.

Tarr, M. J., & Bülthoff, H. H. (1998). Image-based object recognition in man, monkey and machine. In M. J. Tarr & H. H. Bülthoff (Eds), *Object recognition in man, monkey, and machine* (pp. 1-20). Cambridge, MA: MIT Press.

Tarr, M. J., & Bülthoff, H. H. (1999). *Object recognition in man, monkey and machine*. Cambridge, Massachusetts: MIT Press.

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape-recognition. *Cognitive Psychology, 21(2)*, 233-282.

Ullman, S. (1998). Three-dimensional object recognition based on the combination of views. In M. Tarr & H. H. Bülthoff H (Eds.), *Object recognition in man, monkey and machine* (pp. 21-44). England: MIT Press.

Verfaillie, K. (1992). Variant points of view on viewpoint invariance. *Canadian Journal of Psychology, 46*, 215–235.

Wang, G., Obama, S., Yamashita, W., Sugihara, T., & Tanaka, K. (2005). Prior experience of rotation is not required for recognizing objects seen from different angles. *Nature Neuroscience, 8(12)*, 1768–1775.

Wolf, O. T., & Kirschbaum, C. (2002). Endogenous estradiol and testosterone levels are associated with cognitive performance in older women and men. *Hormones and Behavior, 41,* 259-266.

Wolfe, J. M. (1994). Visual search in continuous, naturalistic stimuli. *Vision Research, 34,* 1187-1195.

Wolfe, J. M., Oliva, A., Horowitz, T. S., Butcher, S. J., & Bompas A. (2002). Segmentation of objects from backgrounds in visual search tasks. *Vision Research, 42,* 2985-3004.

Yang-Wallentin, F., Schmidt, P., Davidov, E., & Bamberg, S. (2004). Is there any interaction effect between intention and perceived behavioral control? *Methods of Psychological Research Online, 8*, 127-157.

Zec, R. F. (1995). The neuropsychology of aging. *Experimental Gerontology*, *30(3/4)*, 431-442.

12. DANKSAGUNG

Die Fertigstellung dieser Arbeit war aus mehreren Gesichtspunkten eine aussergewöhnliche und sehr schöne Erfahrung. Insbesondere möchte ich mich bei denjenigen Personen bedanken, welche mich auf diesem Weg begleitet, unterstützt, gestärkt und auch gefördert haben. Bei Herr Prof. Dr. Adrian Schwaninger möchte ich mich für seine Betreuung bedanken. Er hat mir die Freude am wissenschaftlichen Arbeiten auf den Weg mitgegeben und mir gelernt wie wissenschaftliche Experimente geplant, durchgeführt, ausgewertet und anschliessend publiziert werden. Ausserdem habe ich erfahren dürfen, dass wissenschaftliche Erkenntnisse eben auch in der Arbeitswelt zur Anwendung gelangen können.

Ebenfalls möchte ich Herr Prof. Dr. Wolfgang Marx, meinem Erstreferenten, ganz besonders danken. Er hat ein Arbeitsumfeld geschaffen, welches mir erlaubte meine eigenen Ideen zu verwirklichen. Ebenfalls möchte ich mich bei meinem zweiten Referenten Herr Prof. Dr. Mike Martin ganz herzlich für seine fachliche Unterstützung bei den Altersstudien bedanken. Er hat bei beiden Studien zu den Alterseffekten massgeblich mitgearbeitet und durch seine wertvollen Inputs mein Interesse an der Gerontopsychologie geweckt.

Ausserdem möchte ich mehreren Personen der Visual Cognition Research Group (VICOREG) meinen Dank aussprechen. Ganz besonders möchte ich Frau Dr. Franziska Hofer danken, eine gute Fee, welche mir nicht nur fachlich, sondern auch persönlich immer wieder zur Seite stand. Frau lic. phil. Saksia Koller und Herr lic. phil. Stefan Michel möchte ich für die Zusammenarbeit bei der Trainingsstudie und Frau lic. phil. Judith Riegelnig für den wertvollen Input zur Altersstudie danken.

Die zahlreichen Studien wären ohne die grossartige Zusammenarbeit mit der Kantonspolizei Zürich, Kontrollabteilung der Flughafenpolizei, nicht möglich gewesen. Insbesondere möchte ich mich bei Herr Werner Wüest (dem ehemaligen Chef der Kontrollabteilung der Flughafenpolizei), welcher dieses Projekt von Anfang an mitbegleitet hat und uns stets die notwendigen Mittel zur Verfügung stellte, bedanken. Ebenso danke ich auch den zahlreichen weiteren Personen, welche freiwillig an den Experimenten teilgenommen haben.

Einen ganz besonderer Dank geht an meine Eltern und meine Freunde, welche mich während der ganzen Zeit emotional unterstützt haben. Vor allem meinen Eltern und Herr Ivo Vigorelli möchte ich ein ganz herzliches Dankeschön für ihre Unterstützung aussprechen.

13. CURRICULUM VITAE

WORKPLACE AND PERSONAL DATA

Name	Hardmeier Diana
Workplace	University of Zurich Department of Psychology General Psychology (Cognition) Visual Cognition Research Group (VICOREG) Binzmühlestrasse 14/22 CH-8050 Zurich AND Zurich State Police Airport Division, Security Control Branch CH-8058 Zurich-Airport
Phone	UNI: +41 44 635 74 61 Airport: +41 44 655 59 15
Email	UNI: d.hardmeier@psychologie.uzh.ch Airport: hana@kapo.zh.ch
Nationality	Switzerland
Date and Place of birth:	01.08.1976, Switzerland

WORKING EXPERIENCE

Since June 2006	Research associate quality management Zurich state police (50%), airport division, security control branch
Since June 2005	Research assistant at the University of Zurich (50%), Department of Psychology, General Psychology (Cognition)

Since Aug 2004	Project Manager Testing, Visual Cognition Research Group, Department of Psychology, University of Zurich
Since March 1998	Freelance Flight Attendant and Instructor for Crossair/Swiss International Air Lines Ldt., Zurich
March 2002 – Dec 2004	Undergraduate research assistant at University of Zurich, Department of Psychology, General Psychology (Cognition)
March 2004 – May 2004	Internship at Zurich Airport, Airport Police, 8058 Zurich
Aug 2003 – Sep 2003	Internship at AssessmentCoaching, Hueffer & Partner, AssessmentCoaching, Zug

EDUCATION

Since June 2005	Doctoral student at University of Zurich, Department of Psychology, General Psychology (Cognition)
1998 – 2005	Studies of Psychology, Business Management and Journalism, University of Zurich (Lizenziat: June 2005)
1993 – 1998	High school, Kantonsschule Stadelhofen (Typus D)
1989 – 1993	Secondary school Stäfa (ZH) and Jona (SG)
1983 – 1989	Primary school Stäfa (ZH)

FURTHER EDUCATION

Aug 2007	Summer School on Advanced Methods in the Social Sciences: Structural Equation Modeling (SEM)
Since 2006	Studies of the Teaching Skills Program, University of Zurich
Oct 2005 – Feb 2006	English conversation course "Socializing with Colleagues & Visitors" at University/ETH Zurich